2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

Classification of Clothing using Interactive Perception

Bryan Willimon, Stan Birchfield, and Ian Walker
Department of Electrical and Computer Engineering
Clemson University, Clemson, SC 29634
{rwillim,stb}@clemson.edu, ianw@ces.clemson.edu

Abstract— **We present a system for automatically extracting and classifying items in a pile of laundry. Using only visual sensors, the robot identifies and extracts items sequentially from the pile. When an item has been removed and isolated, a model is captured of the shape and appearance of the object, which is then compared against a database of known items. The classification procedure relies upon silhouettes, edges, and other low-level image measurements of the articles of clothing. The contributions of this paper are a novel method for extracting articles of clothing from a pile of laundry and a novel method of classifying clothing using interactive perception. Experiments demonstrate the ability of the system to efficiently classify and label into one of six categories (pants, shorts, short-sleeve shirt, long-sleeve shirt, socks, or underwear). These results show that, on average, classification rates using robot interaction are 59% higher than those that do not use interaction.**

Fig. 1. The proposed setting for interactive perception. A robotic arm interacts with a pile of unknown objects (laundry) to isolate the individual items one at a time and to learn each object's characteristics. The system then classifies the item automatically by comparing the appearance and shape of the object with those of a learned database. Sensory information is provided by a pair of overhead stereo cameras and a side-facing camera (not shown).

I. INTRODUCTION

Laundry is a daily routine throughout the world for people of all walks of life. While this routine was changed significantly more than a century ago when the tasks of washing and drying were automated by modern-day appliances, the remaining tasks of sorting and folding clothes are still performed manually even today as they have been for thousands of years. Nevertheless, with the recent explosion of interest and development in household service robots, there is a realistic possibility that these remaining parts of the laundry process may be automated within the coming generations.

Manipulating and interacting with non-rigid objects remains a largely unsolved problem for robotics, with most research focused upon rigid objects [1]. However, some work in this area is beginning to emerge [2], such as motion planning algorithms for deformable linear objects (DLOs) like ropes, cables, and sutures [3] [4]; or Probabilistic RoadMap (PRM) planners to plan paths for a flexible surface patch [5] or deformable object [6]; or recursive learning approaches [7] or range sensors to sense and model deformable surfaces [8] [9]. Other research has focused upon the problem of manipulating fabric, particularly for textile applications. For example, accurate CAD-like models of textiles allow the computation of tangential stresses for laying, folding, and flattening textiles using vision and force sensing [10] [11] [12].

The problem of automating laundry in particular has received increasing attention lately. Researchers focused upon service robot applications have developed systems for grasping clothes [13], [14], folding clothes [15] [16], and tracing edges [17] [18]. Keio University's "Foldy" mobile robot has demonstrated the ability to fold a shirt based on high-level user input.[1] Similar research at other universities aimed at folding manipulation has made progress on folding origami [19], T-shirts,[2] and towels [18]. Others have developed cardboard machines to fold T-shirts.[3] These existing systems all assume that an individual article of clothing has been isolated and laid flat on a surface prior to manipulation.

In this paper, we present a system based on interactive perception for automatically sorting laundry. See Figure 1. In contrast to previous work, our system operates on an unorganized, unflattened pile of laundry for the purpose of isolating and classifying each individual item. The first task, namely isolating an individual article of clothing, involves identifying and extracting an item from the pile, one at a time, without disturbing the rest of the clothes in the pile. The second task, namely classifying an item, requires using visual-based shape and appearance information to classify the item into one of several prespecified categories (pants, shorts, short-sleeve shirt, long-sleeve shirt, socks, or underwear). Our approach relies upon a combination of graph-based image segmentation, stereo matching, and low-level image comparisons in order to accomplish these objectives. The proposed method can be seen as a particular application of the paradigm of *interactive perception*, also known as *manipulation-guided sensing*, in which the manipulation is used to guide the sensing in order to gather information not obtainable through passive sensing alone [20] [21] [22] [23] [24]. In other words, deliberate actions change the state of the world in a way that simplifies perception and consequently

[1] http://inventorspot.com/articles/laundryfolding_robot_learns_job_34327
[2] http://www.cs.dartmouth.edu/~robotics/movies/mpb-movie-09-shirt-folding.mov
[3] http://www.metacafe.com/watch/1165247/clothes_folding_machine

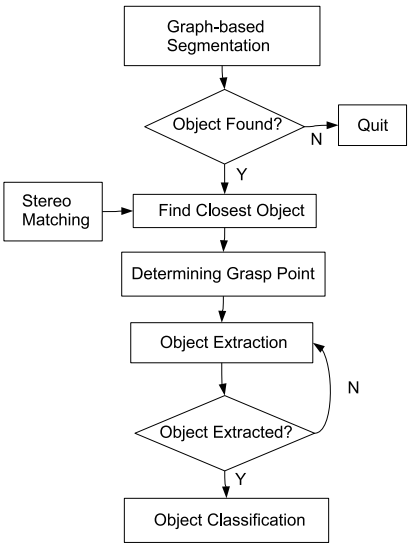

Fig. 2. Overview of our system for vision-guided extraction of items from a pile of laundry.

future interactions.

II. APPROACH

Our laundry-handling system involves two parts: isolating items from a pile of laundry and classifying isolated items. These two parts are described in detail in this section.

A. Isolating items

To isolate an item from the pile, an overhead image is first segmented, and the closest foreground segment (measured using stereo disparity) is selected. Chamfering is used to determine the grasp point which is then used to extract the item from the pile in an fully automatic way using interactive perception. An overview of the process is shown in Figure 2.

1) Graph-based segmentation: The first step is to segment one of the overhead images into different regions. We use Felzenswalb and Huttenlocher's graph-based segmentation algorithm [25] because of its straightforward implementation, effective results, and efficient computation. This algorithm uses a variation of Kruskal's minimum-spanning-tree algorithm to iteratively cluster pixels in decreasing order of their similarity in appearance. An adaptive estimate of the internal similarity of the clusters is used to determine whether to continue clustering. Figure 3 shows the results of graph-based segmentation on an example 320×240 RGB color image taken in our lab using the default value for the scale parameter ($k = 500$). As can be seen, the segmentation provides a reasonable representation of the layout of the items in the pile. From this result, we determine the foreground regions as those that do not touch the boundary of the image.

2) Stereo matching: Since our goal is to remove a single piece of clothing without disturbing the remaining items in the pile, the next step in the process is to determine which

Fig. 3. LEFT: An image taken by one of the overhead cameras in our setup. RIGHT: The results of applying the graph-based segmentation algorithm. Despite the over-segmentation, the results provide a sufficient representation for grasping an article of clothing.

item is on top of the pile. While there are many monocular image cues that can provide a hint as to which object is on top, such as the size of the object, its concavity, T-junctions, and so forth, we rely upon stereo matching due to its efficiency, ease of implementation, and robustness. Stereo matching is the process in visual perception leading to the sensation of depth from two slightly different projections of an environment [26]. With rectified cameras, the difference in image coordinates between two corresponding points (the horizontal disparity) is inversely proportional to the distance from the camera to the point. We implemented an 11×11 window-based sum-of-absolute differences (SAD) stereo algorithm for its computational efficiency, utilizing MMX/SSE2 SIMD operations and a running sum sliding window to increase the speed of computation.

Due to misalignment of the cameras, reflections in the scene (non-Lambertian surfaces), and occlusion, the resulting disparity image is noisy. To reduce the effects of this noise, we employ a left-right consistency check [27] to retain only those disparities that are consistent in both directions. Photometric inconsistency between the cameras is handled by converting to grayscale, followed by adjusting the gain of one image to match the other. After computing the disparities in this manner, the relative height of each segmented region is determined by the average disparity of all the pixels in the region. Among the foreground regions exceeding a minimum size (0.04% of the image), the one with the largest average disparity is then estimated as the item on top of the pile.

3) Determining the grasp point: Once the top item has been determined, the next step is to determine the grasp point of the item. We cannot use the approach of Saxena et al. [28] due to the use of non-rigid and irregularly shaped objects. Instead, we calculate the 2D grasp point as the geometric center of the object, defined as the location whose distance to the region boundary is maximum. This point, which can be computed efficiently using chamfering [29], is much more reliable than the centroid of the region, particularly when the region contains concavities which is not uncommon in our scenarios. Figure 5 shows an example of the grasp point found by the maximum chamfer distance for an article of clothing. Note that only one grasp point is found due to the limitation of using a single robotic manipulator.

4) Extracting the item: Once the grasp point has been found, the robot arm is moved over the pile of laundry so that

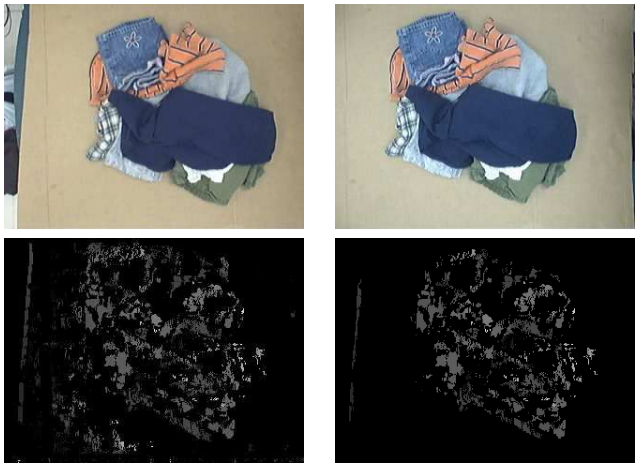

Fig. 4. TOP: A stereo pair of images taken by the overhead cameras in our setup, showing the large amount of photometric inconsistency. BOTTOM: The disparity image obtained by SAD matching with the left-right disparity check (left), and the result after masking with the foreground and removing small regions (right).

Fig. 5. LEFT: The binary region associated with an article of clothing (orange shirt), with the grasp point (red dot) computed as the location that maximizes the chamfer distance. RIGHT: The chamfer distance of each interior point to the clothing boundary.

the end effector is positioned above the grasp point. The arm then engages in a procedure that we refer to as *bobbing*. The arm is lowered to just above the estimated height of the item, then end effector is closed, the arm is raised and moved to the side. During this process the presence of the arm occludes the scene, making the images of the overhead cameras uninformative. Therefore, after completion of the process, the images before and after are compared to determine whether the desired item was extracted from the pile. Simple frame differencing with a threshold is used to make this decision. If it is determined that no item was extracted, then the procedure is repeated, this time with the end effector reaching down further. The process is repeated successively at increasing distances until either the item is removed or the end effector reaches the minimum height above the table (to avoid collision). If all attempts are unsuccessful, then the robot arm is returned to the home position, and the entire process begins again; in our experiments the robot was always successful using no more than two bobbing procedures. The entire procedure for extracting a single item is repeated until no more objects remain in the pile, which is determined via a threshold on the minimum size of the foreground regions in the segmented image (0.5% of the image).

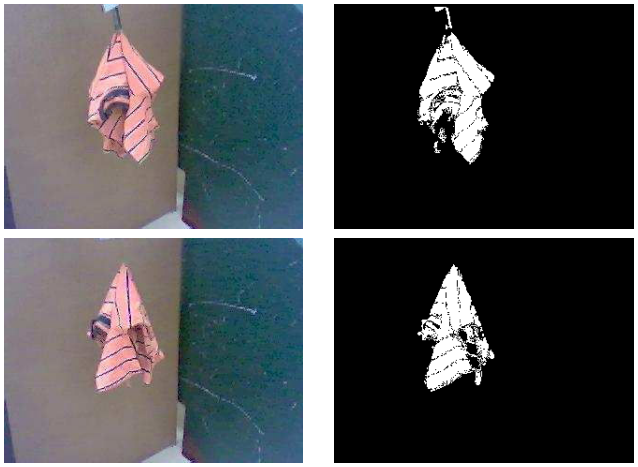

Fig. 6. The front (top) and side (bottom) views of an isolated article of clothing to be classified (orange shirt). In each case, the original image is shown on the left, while the binary silhouette is shown on the right.

Note that this bobbing procedure is a form of interactive perception. Because our sensors are not accurate enough to precisely compute the distance to the object, and because our gripper is not guaranteed to be oriented in the correct direction, it is virtually impossible to ensure success on the very first try. Moreover, our particular robot is not equipped with a force sensor, thereby increasing the difficulty of sensing the environment. To overcome this limitation in sensing, an interactive sensing (yet fully automatic) approach is adopted in which repeated interactions with the environment are used to simplify the problem of sensing.

B. Classifying items

Once the robot removes and isolates an article of clothing from the pile, the arm lifts and swings so that the article hangs freely without touching the table or ground. This open area is monitored by a third side-facing camera that captures both a "frontal view" and a "side view" image of the article, using the robot to rotate 90 degrees about the vertical axis between images. These two views are subtracted from a background image to obtain two binary silhouettes of the clothing. Figure 6 illustrates an example of the two views of an isolated article of clothing along with its binary silhouettes. Note that the terms "frontal" and "side" are arbitrary designations, since the robot grasps each object somewhat at random. What is important is that these two silhouttes represent the object's shape from two orthogonal directions (and, by symmetry, from the other two orthogonal directions by a mirror flip).

Features are extracted from the frontal and side images of the article of clothing in order to compare with other previously labeled images of clothing. Let I_Q be one of these two query images (either frontal or side), and let I_D be an image in the database. These images are compared using four

different features to yield a match score:

$$\Phi(I_Q, I_D) = \sum_{i=1}^{N} w_i \cdot \frac{1}{m_i} f_i(I_Q, I_D), \quad (1)$$

where $N = 4$ is the number of features, w_i is the weighting associated with each feature, and the features are given by

- $f_1(I_Q, I_D) = |a_Q - a_D|$, the absolute difference in area between the two silhouettes, where a_Q is the number of pixels in the query binary silhouette, and similarly for a_D;
- $f_2(I_Q, I_D) = |e_Q - e_D|$, the absolute difference in eccentricity between the two silhouettes, where $0 \leq e_Q \leq 1$ is the eccentricity of the query binary silhouette, and similarly for e_D;
- $f_3(I_Q, I_D) = H(BE_Q, BE_D)$, the Hausdorff distance between the edges of the two binary silhouettes;
- $f_4(I_Q, I_D) = H(CE_Q, CE_D)$, the Hausdorff distance between the Canny edges of the original grayscale images.

To ensure proper weighting, each value m_i is the 95th percentile of f_i among all the images in the database (robust maximum).

Although color information would be a helpful cue for identifying particular items of clothing, it was not used in this work because color varies so widely within clothing categories. After calculating the match scores, the nearest neighbor algorithm (1-NN) was used to assign a category to the article of clothing. The 1-NN algorithm finds the category associated with the silhouette in the database whose match score to the test silhouette is minimum.

As will be seen in the experimental results, the accuracy of the above procedure is rather poor. This is due to the impoverished sensing that occurs when an article of clothing hangs freely from a single grasp point. To overcome this limitation, we use interactive perception. The process of capturing front and side views of an item is repeated ten times. In each iteration, the robot drops the item on the table, and the item is extracted again in a manner similar to that described above. The randomness of the dropping results in a new grasp point that, in general, bears little relationship to previous grasp points. These multiple grasp points provide the system with multiple front and side views of the article of clothing, thereby greatly increasing the chance of accurately classification.

III. Experimental Results

The proposed approach was applied in a laundry scenario to test its ability to perform practical interactive perception using a PUMA 500 robotic arm. In each experiment, a pile of laundry rested upon a flat, uniform background. The articles of clothing in the pile consisted of five short sleeve shirts, five long sleeve shirts, five pairs of pants (trousers), five shorts, five socks, and five pairs of underwear. The articles themselves, their type, and their number were unknown to the system beforehand.[4]

[4]Due to the physical limitations of the workspace, we used children's clothing. This introduces an overall scaling factor to the sensory data.

A. Extraction Experiment

Figure 7 shows the results of the system after different steps of the extraction and isolation procedure. The system removes items from the pile one at a time by identifying a grasp point in the closest region, then deploying the arm to that location, bobbing for the item, and moving the item to another part of the workspace. Once the item has been extracted and isolated, the system interacts with it by rotating about a vertical axis in order to gain both front and side views of the item using the side-facing camera. Two 2D silhouettes of the item are created for each grasp point; by repeating this process ten times, 20 silhouettes are obtained for each item (ten front views and ten side views). The procedure successfully removed all items from the pile one at a time.

B. Classification Experiments

With 6 categories, 5 items per category, and 20 images per item, the database collected by the extraction / isolation procedure consists of 600 images. This database was labeled in a supervised learning manner so that the corresponding category of each image was known. Eight tests were used to compare the training and test images:

test name	w_1	w_2	w_3	w_4
Area	1	0	0	0
Eccentricity	0	1	0	0
Binary Edges	0	0	1	0
Canny Edges	0	0	0	1
Combo A	1	1	0	0
Combo B	1	1	0	1
Combo C	0	1	0	1
Combo D	1	0	0	1

We conducted two experiments: leave-one-out classification and train-and-test classification. In leave-one-out classification, each of the 600 images was compared with the remaining 599 images in the database. If the nearest neighbor among these 599 was in the same category as the image, then the classification was considered a success, otherwise a failure. Results for all 100 images for each category were combined to yield a classification rate for that category, and the procedure was repeated for all eight tests. The results, shown in Figure 8, reveal that most categories were either classified, on average, well (near or above 50%) or poorly (near or below 30%), with the best results achieved by the Combo A and Combo B tests.

In train-and-test classification, three articles of clothing from each category were selected for the training set and the remaining two articles were used for the test set. Therefore, each image was compared with the 360 images in the training set, and the category of the nearest neighbor among these images was compared with the category of the image to determine success or failure. Results for all 40 test images for each category were combined to yield a classification rate for that category, and the procedure was repeated for all eight tests. The results, shown in Figure 9, show results that are significantly worse than those of the leave-one-out experiment, due to the reduced amount of data used for

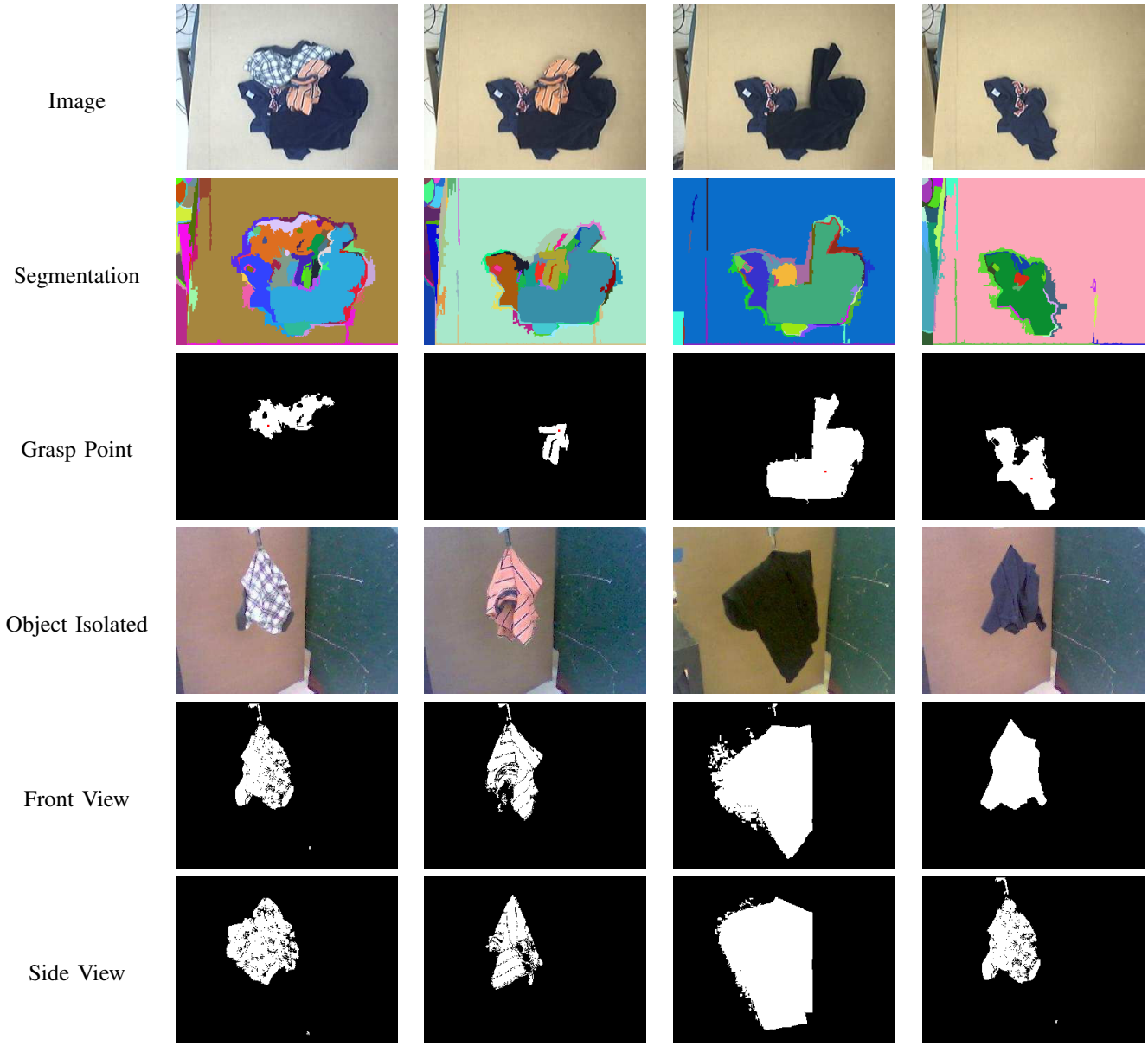

Fig. 7. The extraction and isolation process. From top to bottom: The image taken by one of the downward-facing stereo cameras, the result of graph-based segmentation, the object found along with its grasp point (red dot), the image taken by the side-facing camera, and the binary silhouettes of the front and side views of the isolated object. Time flows from left to right, showing the progress as each individual article of clothing is removed from the pile and examined.

training. We should emphasize that classifying an unknown article of clothing (hanging from a single, arbitrary grasp point) from a single image is extremely difficult even for a human viewer.

C. Interaction vs. Non-interaction

One of the goals of this work was to determine the usefulness of interactive perception in a clothing classification context. The process of interacting with each article of clothing provided the system with multiple views using various grasp locations, allowing the system to collect 20 total images of each object. Therefore, in the next experiment we compared features from all 20 images of each object with the remaining images in the database. The procedure was as follows. The query article of clothing was compared with all articles in the database, and the category of the closest matching article was considered to be the category of the query article. Two articles were compared by examining the 400 match scores between their pairs of images (20 images per article). For each of the 20 images of the query article, the 1-NN among the other 20 images was found; these 20 match scores were then added to yield the distance between the two articles. This procedure corresponds to a manipulator picking up and dropping an object multiple times to get multiple views of it in order to better classify it. The results, shown in Figure 10, are indeed significantly higher than those obtained without interactive perception. In fact, Combo A achieved

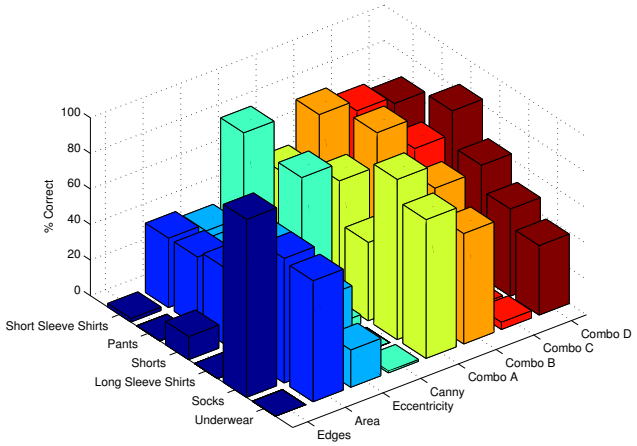

Fig. 8. Leave-one-out classification results for all six categories using eight different tests.

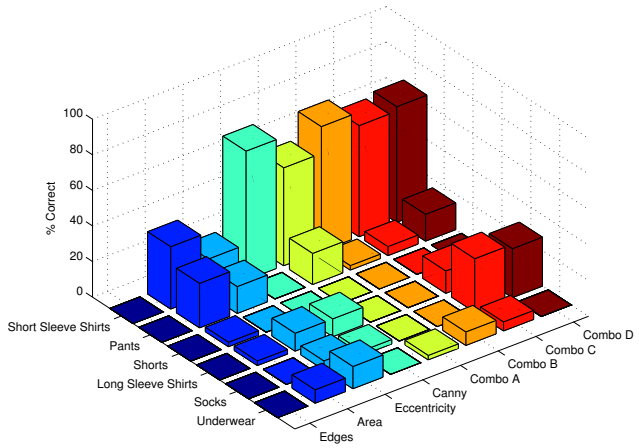

Fig. 9. Train-and-test classification results for all six categories using eight different tests.

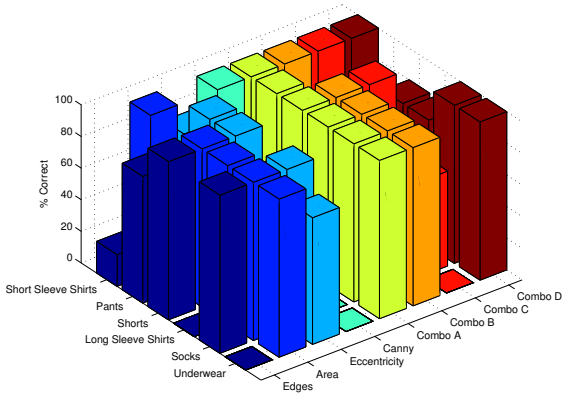

Fig. 10. Classification results for all 30 objects using all 20 images for comparison.

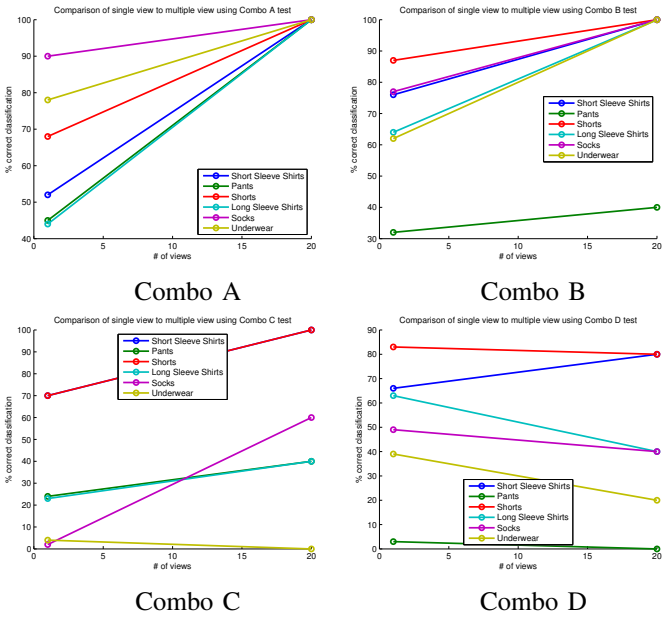

Fig. 11. Classification results for all 30 objects using all 20 images for comparison using the combination tests.

100% classification on all categories using this method.

For comparison, the classification rates for all categories for the four different combination tests are shown in Figure 11, illustrating the difference between using a single view versus using all 20 views of an object. For Combo A, the average classification rate using a single image is 62.83%, while the average classification rate using all 20 images is 100%. These results show that, on average, classification rates using robot interaction are 59% higher than those that do not use interaction.

IV. Conclusion

We have proposed an approach to interactive perception in which a pile of laundry is sifted by an autonomous robot system in order to classify and label each item. The algorithm is shown empirically to provide a practical way to extract items out of a cluttered area one at a time with minimal disturbance to the other objects. This system uses a single color camera to segment the various items within the scene and calculate a location for the robot arm to grasp the next object for extraction. A stereo pair of cameras is used to calculate depth in order to identify the closest item (i.e., the object on top of the pile). The system takes approximately one minute per article of clothing for the combined steps of location, grasp point identification, extraction, and classification of the item.

Several directions for future research can be explored in this novel area of interactive perception in cluttered environments. First, additional interaction and labeling techniques could be used to improve the ability of the system to determine which characteristics of an object make it distinguishable from other objects. Secondly, the learning process could be improved by enabling the system to learn about categories of clothing for which it has not been trained. A third improvement could be to identify the owner of an article of clothing which has been seen before, in order to stack the articles according to the person expected to wear them. Another modification would be to address the

problem of classifying bright versus dark clothes prior to loading the laundry into the washing machine. Finally, the modeling of the items could incorporate 3D information in order to provide a more accurate representation of the front and side views that describe the overall shape of the object. We believe that all of these areas are promising future extensions of this research.

V. Acknowledgments

This research was supported by the U.S. National Science Foundation under grants IIS-1017007, IIS-0844954, and IIS-0904116.

References

[1] D. Kragic, M. Björkman, H. I. Christensen, and J.-O. Eklundh, "Vision for robotic object manipulation in domestic settings," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 85–100, Jul. 2005.

[2] D. Henrich and H. Worn, Eds., *Robot Manipulation of Deformable Objects: Advanced Manufacturing*. New York: Springer Verlag, 2000.

[3] M. Saha and P. Isto, "Motion planning for robotic manipulation of deformable linear objects," in *Proceedings of the International Conference on Robotics and Automation*, May 2006, pp. 2478–2484.

[4] M. Moll and L. E. Kavraki, "Path planning for deformable linear objects," *IEEE Transactions on Robotics*, vol. 22, no. 4, pp. 625–636, Aug. 2006.

[5] C. Holleman, L. E. Kavraki, and J. Warren, "Planning paths for a flexible surface patch," in *Proceedings of the International Conference on Robotics and Automation*, May 1998, pp. 21–26.

[6] O. B. Bayazit, J.-M. Lien, and N. M. Amato, "Probabilistic roadmap motion planning for deformable objects," in *Proceedings of the International Conference on Robotics and Automation*, May 2002, pp. 2126– 2133.

[7] A. Howard and G. Bekey, "Recursive learning for deformable object manipulation," in *Proceedings of the 8th International Conference on Advanced Robotics (ICAR)*, Jul. 1997, pp. 939–944.

[8] P. Fong, "Sensing, acquisition, and interactive playback of data-based models for elastic deformable objects," *International Journal of Robotics Research*, vol. 28, no. 5, pp. 622–629, May 2009.

[9] Y. Kita, F. Saito, and N. Kita, "A deformable model driven visual method for handling clothes," in *International Conference on Robotics and Automation*, 2004.

[10] M. Bordegoni, G. Frugoli, and C. Rizzi, "Direct interaction with flexible material models," in *Proceedings of the Seventh International Conference on Human-Computer Interaction (HCI)*, Aug. 1997, pp. 387–390.

[11] M. Fontana, C. Rizzi, and U. Cugini, "3D virtual apparel design for industrial applications," *Computer-Aided Design*, vol. 37, no. 6, pp. 609–622, 2005.

[12] K. Paraschidis, N. Fahantidis, V. Petridis, Z. Doulgeri, L. Petrou, and G. Hasapis, "A robotic system for handling textile and non rigid flat materials," *Computers in Industry*, vol. 26, pp. 303–313, 1995.

[13] P. Gibbons, P. Culverhouse, and G. Bugmann, "Visual identification of grasp locations on clothing for a personal robot," in *Conference Towards Autonomous Robotic Systems (TAROS)*, Aug. 2009, pp. 78–81.

[14] K. Yamakazi and M. Inaba, "A cloth detection method based on image wrinkle feature for daily assistive robots," in *IAPR Conference on Machine Vision Applications*, 2009, pp. 366–369.

[15] K. Salleh, H. Seki, Y. Kamiya, and M. Hikizu, "Inchworm robot grippers for clothes manipulation," *Artificial Life and Robotics*, vol. 12, no. 1–2, pp. 142–147, 2008.

[16] F. Osawa, H. Seki, and Y. Kamiya, "Clothes folding task by tool-using robot," *Journal of Robotics and Mechatronics*, vol. 18, no. 5, pp. 618–625, 2006.

[17] K. Salleh, H. Seki, Y. Kamiya, and M. Hikizu, "Tracing manipulation of deformable objects using robot grippers with roller fingertips," in *International Joint Conference on SICE-ICASE*, Oct. 2006, pp. 5882–5887.

[18] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *International Conference on Robotics and Automation (ICRA)*, 2010.

[19] D. J. Balkcom and M. T. Mason, "Robotic origami folding," *International Journal of Robotics Research*, vol. 27, no. 5, pp. 613–627, May 2008.

[20] D. Katz and O. Brock, "Manipulating articulated objects with interactive perception," in *Proceedings of the International Conference on Robotics and Automation*, May 2008, pp. 272–277.

[21] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *International Conference on Robotics and Automation (ICRA)*, 2009, pp. 1377–1382.

[22] P. Fitzpatrick, "First contact: An active vision approach to segmentation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2003.

[23] B. Willimon, S. Birchfield, and I. Walker, "Rigid and non-rigid classification using interactive perception," in *International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[24] R. B. Willimon, "Interactive perception for cluttered environments," Master's thesis, Clemson University, Dec. 2009.

[25] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[26] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[27] P. Fua, "Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, 1991, pp. 1292–1298.

[28] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *International Journal of Robotics Research*, vol. 27, pp. 157–173, Feb. 2008.

[29] G. Borgefors, "Distance transformations in digital images," *Computer Vision, Graphics, and Image Processing*, vol. 34, no. 3, pp. 344–371, 1986.