2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Cloth Grasp Point Detection based on Multiple-View Geometric Cues with Application to Robotic Towel Folding

Jeremy Maitin-Shepard, Marco Cusumano-Towner, Jinna Lei and Pieter Abbeel

Abstract— **We present a novel vision-based grasp point detection algorithm that can reliably detect the corners of a piece of cloth, using only geometric cues that are robust to variation in texture. Furthermore, we demonstrate the effectiveness of our algorithm in the context of folding a towel using a general-purpose two-armed mobile robotic platform without the use of specialized end-effectors or tools. The robot begins by picking up a randomly dropped towel from a table, goes through a sequence of vision-based re-grasps and manipulations—partially in the air, partially on the table—and finally stacks the folded towel in a target location. The reliability and robustness of our algorithm enables for the first time a robot with general purpose manipulators to reliably and fully-autonomously fold previously unseen towels, demonstrating success on all 50 out of 50 single-towel trials as well as on a pile of 5 towels.**

I. INTRODUCTION

In highly structured settings, modern-day robots can be scripted to perform a wide variety of tasks with mind-boggling precision and repeatibility. However, outside of carefully controlled settings, robotic capabilities are much more limited. Indeed, the ability to even merely grasp a modest variety of previously unseen rigid objects in real-world cluttered environments is considered a highly non-trivial task [26], [13], [2].

Handling of non-rigid materials, such as clothing, poses additional challenges: they tend to have significantly higher dimensional configuration spaces and, resultingly, a large variety of visual appearances. Aside from the environmental clutter, the object itself could occlude otherwise desirable grasp points. Moreover, when manipulating deformable objects, the grasp point will typically affect the configuration of the object—hence when deciding upon grasp points one needs to account for both whether the robot arm can reach them in an appropriate way and whether that particular grasp would result in an acceptable configuration of the deformable object. This results in significant additional challenges in grasp point detection and selection.

The challenges involved in manipulation of deformable objects are greatly reflected in the state of the art in robotic laundry folding. Indeed, the current state of the art is far from enabling general purpose manipulators to fully automate the task of laundry folding. In fact, even including work that uses assisting tools such as flip-folds, thus far no comprehensive success story has been reported for the complete end-to-end task of reliably picking up a laundry item and folding it.

The authors are with the Department of Electrical Engineering and Computer Sciences, UC Berkeley, CA 94720, U.S.A. Email: {jbms,pabbeel}@cs.berkeley.edu, {marcoct,jinna.lei}@berkeley.edu.

In this paper we present an algorithm which addresses one of the critical challenges in automating laundry folding: grasp point detection. In particular we focus on one category of grasp points which is of general interest: detection of corners of an article of clothing. For many articles of clothing, the ability to grasp certain corners is a key enabler towards spreading out and then folding. For the particular task of folding a towel, being able to grasp the corners reliably provides many natural paths to satisfactory solutions which are executable by modern-day robots. Perhaps not surprisingly, going from a towel randomly placed on a surface area to holding up the towel by two of its corners is a missing link in the prior research on automating the towel folding task.

We leverage our algorithm's capability of reliably detecting the corners of a towel to develop the first system that uses two generic robotic arms (and a pair of cameras) and is able to reliably pick up and fold a towel.

Videos of our experimental results are available at:

http://rll.eecs.berkeley.edu/pr/icra10

II. RELATED WORK

We refer the reader to existing surveys [4], [27] for a more comprehensive literature review on the broad domain of grasping research. We focus on the areas that are most closely related to our work: Vision-based grasping in real-world environments, and prior work towards robotic laundry folding.

There have been some encouraging results on grasping rigid objects using vision. The earliest work was mostly limited to grasping 2-D planar objects. Examples include, but are not limited to, [6], [19], [5], [12]. More recently, there has also been substantial progress on vision based grasping for objects with 3-D structure in cluttered environments. Saxena and collaborators [26] have proposed algorithms for grasping previously unseen objects in cluttered environments. They use supervised machine learning to learn good grasp point candidates from labeled training data. Platt and collaborators [23], [24] leverage learning for vision-based grasping in varying contexts, including bagging groceries. Berenson and collaborators [2], [3] proposed algorithms for path planning for vision-based grasping in cluttered environments.

From the application angle, the prior work on robotic laundry manipulation and folding is most closely related to our work. To the best of our knowledge none of the prior work—including the prior work which uses special-purpose end-effectors—reports successful completion rates or even successful completions for the full end-to-end task

for picking up an arbitrarily placed towel or clothing article and bringing it into a nicely folded end-state.

Paraschidis and collaborators [7] describe the isolated executions of grasping a layed-out material, folding a layed-out material, laying out a piece of material that was already being held, and flattening wrinkles.

Some of the prior work describes robots using tools and the design of special purpose end-effectors as a step towards laundry folding. For example, Osawa and collaborators [21] developed a robot capable of using a "flip-fold" for folding and a plate for straightening out wrinkles. They assume the robot starts the process by already holding two appropriate points of the piece of clothing.

There is also a body of work on recognizing categories of clothing, some of the work includes manipulation to assist in the categorization. For example, Osawa and collaborators [22] and Hamajima and Kakikura [10] present approaches to spread out a piece of clothing using two robot arms and then categorize it.

Salleh and collaborators [25] present an inchworm gripper for tracing the edge of a piece of clothing. The gripper assists in enabling two robotic manipulators to get into the state of holding a piece of cloth by two neighboring corners. They report a 65% success rate on grasping the first corner, identified as the bottommost point of the towel in the image once the towel has been picked up from the table and held up by one of the arms. The overall success rate in holding up the towel by two neighboring corners is 50%. A range of gripper designs is presented in [18].

Yamakazi and Inaba [29] present an algorithm that recognizes wrinkles in images, which in turn enables them to detect clothes laying around. Kita and collaborators [14] fit the geometry of the silhouette of a hanging piece of clothing to the geometry of a mass spring model of the same piece of clothing and are able to infer some 3-D information about the piece of clothing merely from the silhouette.

Balkcom and Mason [1] developed a robot capable of origami.

The work by Kavraki and collaborators [15], [17], by Matsuno and Fukuda [16], and by Gopalakrishnan and Goldberg [9] considers planning for manipulation of deformable objects. While their focus was largely on the planning task and our contribution is primarily in vision based perception aspects, it could be an interesting future direction to also include the clothing article into the motion planner.

III. GRASP POINT DETECTION

A. Overview

Due to the wide range of 3-D deformation possible of even a single article of clothing, and the compounding factor of the wide variation in appearance and material properties across multiple articles of the same type, the geometric properties of *borders* of the cloth, by which we mean actual cuts in or ends of the fabric, are among few other robust local features for grasp point detection. For example, intuitively, the key parts of a t-shirt, namely the sleeves, neck, and lower hem, are well-identified locally as circular interior borders, and may be distinguished by the size; on a towel, the four corners can be locally identified as two exterior borders (of sufficient length) that meet at approximately a right angle (in the surface).

Having established that the border geometry is intuitively quite useful in locating key grasp points of cloth, the question remains of how to actually estimate the border geometry from sensor data. We focus in particular on the case of using images of the cloth for this purpose. Depth discontinuities in the image, identified using, e.g., stereo or foreground-background segmentation, are an obvious cue. Of course, in any particular configuration of the cloth, some borders may not appear as depth discontinuities, or may not appear at all for that matter, but more serious is the problem that folds in the cloth also appear as depth discontinuities. The key distinguishing feature, and which is at the core of our proposed method, is the sharpness of curvature of the cloth.[1]

Depth-discontinuity edges in an image have a key property that is leveraged by our algorithm: given a point $\mathbf{u}$ in an image along an edge with direction $\mathbf{e}$ (which specify a line ℓ in the image) known to correspond to a depth-discontinuity, it must be the case that the 3-D plane P that projects to ℓ is tangent (up to error introduced by the image discretization) to the surface of the object being imaged at the point $\mathbf{p}$ corresponding to $\mathbf{u}$. Thus, if the same point $\mathbf{p}$ (or in general, points within a small region) on the object projects to a depth-discontinuity edge in multiple camera views of the same object (assumed to either be known or estimated using standard techniques in multi-view geometry), then we obtain multiple tangent planes at $\mathbf{p}$ (or for points within a small distance of $\mathbf{p}$), where the maximum angular difference between the normals of these tangent planes provides a lower bound on a measure of curvature at $\mathbf{p}$.

It can be shown that our proposed algorithm computes an approximation to this estimate of curvature based on the amount of observed variation in the tangent plane nearby each point $\mathbf{p}$ that projects to a point on a depth-discontinuity boundary in an image, under the simplifying assumptions that the multiple views correspond to rotations by a fixed incremental angle ϕ of the object about a vertical axis through the center of the object (a true assumption in our experimental setup) and the tangent planes in all of the views correspond to the same plane in the frame of the camera (approximately holds if the optical flow between the images of the point is small).

B. Border Classification Algorithm

Based on the intuition developed in the previous section, we define a filtering algorithm that incrementally processes a sequence of images of an article of clothing in order, incrementally classifying for each image t a subset B_t of the set E_t of points that appear as depth discontinuities in the image as corresponding to actual borders of the cloth.

[1]Of course, a very sharp fold may have identical appearance to a border even from all views, and indeed some borders may in fact be a sewn folded edge, but when the cloth is hanging freely, particularly after having been shaken, such a sharp fold is highly unlikely to remain (unless it is sewn).

Since the algorithm is restricted to classifying points that actually appear as depth discontinuities in the image, a border that happens to be lying flush against another part of the cloth cannot be detected. Furthermore, in order to obtain a very precisely-localized set of depth-discontinuity points at high resolution (important for sufficient sensitivity in distinguishing the curvature of a moderately sharp fold from that of an actual border), a very precise foreground-background segmentation of the input images is computed, and only points on the boundary between foreground and background are considered depth-discontinuities. Thus, only depth discontinuities against the background are considered.[2]

The algorithm incrementally computes for each frame t a score $S_t(\mathbf{u})$ specifying an estimate of the curvature at each depth-discontinuity image edge point $\mathbf{u} = (u, v)$; at all points $\mathbf{u}$ that do not appear as depth-discontinuities in frame t, the score $S_t(\mathbf{u}) = 0$. To track depth-discontinuity points between frames, a dense sub-pixel optical flow map (with associated confidence in $[0, 1]$) from pixels in the current frame to pixels in the previous frame is computed. Interpreting the confidence measure probabilistically, the score $S_t(\mathbf{u})$ is (1 plus) the expected value of the number of consecutive frames prior to t for which the 3-D point $\mathbf{p}$ on the object corresponding to point $\mathbf{u}$ in image frame t has been tracked *while remaining a depth-discontinuity edge*.

Specifically, for the first frame, the score $S_t(\mathbf{u})$ is initialized to 0, and for all subsequent frames is defined by

$$S_t(\mathbf{u}) \equiv \begin{cases} C_t(\mathbf{u}) \cdot S_{t-1}(\mathbf{u} + F_t(\mathbf{u})) + 1 & \text{if } \mathbf{u} \in E'_t; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

E'_t is obtained by dilating the set E_t of depth-discontinuity points for frame t inwards (in the foreground region only) by δ pixels to account for optical flow not being well-defined (and in practice having a very low confidence score) at actual edge points. $F_t(\mathbf{u})$ is the optical flow map that transforms a point $\mathbf{u} = (u, v)$ in the current frame into the corresponding point $\mathbf{u}' = (u', v')$ in the previous frame, and $C_t(\mathbf{u})$ is the corresponding confidence score.[3]

The actual set B_t of border points is computed as the set of edge points $\mathbf{u} \in E_t$ for which the score S_t, averaged over a circular disk (restricted to the foreground only) around $\mathbf{u}$ of radius δ, exceeds a threshold t.

C. Corner Grasp Point Detection

Based on the computed map of border classifications, candidate corner grasp points are selected using a RANSAC-based algorithm that fits corners to the border points [8]. Two border points are considered *compatible* if they are each on the inside of the line corresponding to the other point and the corner they form has an angle between 45° and 110°. Each RANSAC iteration, two compatible border points are randomly sampled from B_t. Because each border point has a corresponding 2-D edge direction (based on the estimated edge direction in the image), two border points, each specifying a 2-D line in the image, are sufficient to specify a corner. Points within d_{inlier} pixels of each of the two rays formed by the corner are considered inliers, and the *size* of each corner side is defined to be the maximum length for which the cumulative gap amount between border points projected back onto the line segment (starting at the corner tip) is less than ℓ_{gap}. The *size* of the corner is defined to be the minimum of the sizes of the two sides. The complete RANSAC procedure consists of many sampling attempts, from which the largest corner with size at least $\ell_{\min}$, if any, is selected. Repeated runs of the complete RANSAC procedure are used to fit as many 2-D corners (with disjoint corresponding inlier sets) as possible in the image.

D. Filtering based on 3-D localization

For each of the 2-D corners found by the RANSAC algorithm, a plane is fit to the high-confidence 3-D points (computed using stereo[4]) corresponding to the image points contained in the corner; the fit is weighted according to a Gaussian kernel centered on the tip. If there is insufficient total weight on the high-confidence points, or more than 10% of the weight lies on points that are more than 0.4 cm from the optimal plane, the corner is rejected.

If the candidate is accepted, the actual grasp point and orientation are chosen according to the desired grasp position and orientation relative to the corner tip as well as workspace constraints of the robot, based on the 3-D triangle given by the plane fit. A chosen grasp position and orientation can be checked for feasibility using a standard motion planner (e.g. PRM-based) that also takes into account the rest of the cloth as an obstacle (by using the stereo-derived point cloud).

IV. APPLICATION TO ROBOTIC TOWEL FOLDING

Integrating as a key component the grasp point detection algorithm described in §III, we implemented a complete procedure for folding and stacking a pile of randomly dropped towels on a table. The procedure, which involves a sequence of vision-based grasps, re-grasps and manipulations, partially in the air, partially on the table, is modeled as a state machine as illustrated in Fig. 2. We used a prototype version of the Willow Garage PR2 robotic platform [28], shown in Fig. 1.

The grasp point for the initial pickup from the pile (Fig. 2a-b) is estimated using a combination of foreground-background segmentation and stereo correspondence applied to the stereo pair on the head to select a central point that can be reached by one of the two arms; a fixed grasp orientation perpendicular to the (known) table is used. If multiple towels

[2]Although this may appear be overly restrictive and lead to a scarcity of detections, it is in fact not nearly as limiting as it may seem: because the focus is on grasp point detection, it is not necessary that a detection be made from every angle, only that it be possible from *some* angle; furthermore, borders that are flush against another part of the cloth (and in a freely hanging configuration, any border that is not against a background is likely to be fairly close to another part of the cloth) would likely require a much more sophisticated motion planning for grasping that takes into account the deformation that would result from contacting the cloth prior to the grasp, in order to avoid accidentally snagging additional parts of the cloth in the process of the grasp. In contrast, borders against the background can typically be grasped using traditional rigid object motion planning.

[3]Because F_t provides a sub-pixel (real-valued) estimate of the flow vector, $S_{t-1}(\mathbf{u} + F_t(\mathbf{u}))$ is evaluated using bilinear interpolation.

[4]Note that the stereo correspondence need only be computed for images where a 2-D corner is detected.

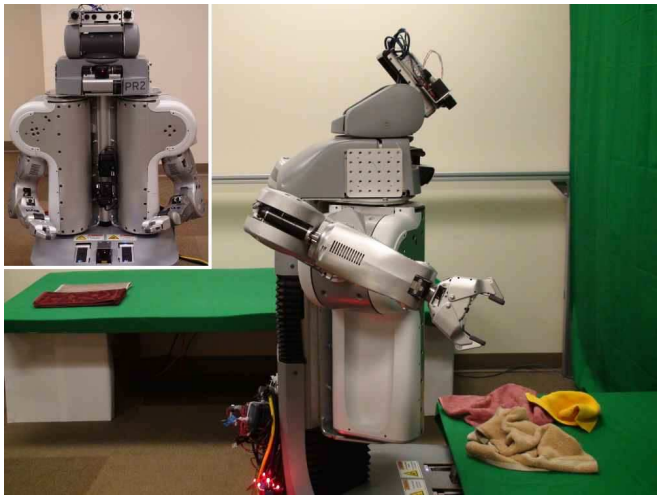

Fig. 1. Our robotic platform (the PR2) consisted of two 7-DOF robotic arms mounted to a 1-DOF vertical-slide torso on a mobile base supporting omni-directional position control. A stereo pair of 640x480 color cameras were mounted to the pan-tilt sensor head, and a stereo pair of high-resolution 3312x4416 (14MP) Canon G10 cameras were mounted between the arms (which provides a superior viewing angle for certain parts of the towel folding procedure). (Other sensors on the robot were not used.) The workspace consists of two known tables: one contains the initial pile of towels, and the other is used for folding and stacking the towels. Note that the PR2 was not designed specifically for this particular task, and no specialized tool or end-effector design was used.

happen to be picked up, all but one is later dropped back in the pile automatically following the first corner grasp.

To provide sufficient motion cues for the main corner grasp point detection algorithm, the robot allows the towel to hang vertically from its gripper and rotates the towel in place in view of the high-resolution stereo pair (Figure 2c-e). At most five attempts of sixteen rotational increments (of 12° each) are made to detect a suitable corner grasp point; each attempt is preceded by a shake maneuver designed to randomize the configuration of the towel. A grasp is attempted as soon as a suitable point that can be reached by a collision-free path (including collisions with the towel) is detected. For simplicity, a blind grasping strategy was used, in which the planned grasp is attempted without additional using visual feedback to track the detected corner. Consequently, it was necessary for the towel to be stationary and in a stable configuration when imaged, which dictated that the towel be rotated slowly and allowed to settle for several seconds prior to capturing each frame.

If all five attempts are exhausted, the towel is dropped back in the pile to allow it to be retried with a different initial pickup. Any detected corner is permitted for the first grasp, while for the second grasp (after the towel is already held from one corner), the bottommost corner is rejected in order to ensure that the towel is held by two adjacent corners. A completely missed or insecure grasp is detected immediately by attempting to pull the towel taut between the two grippers; other types of mis-grasps are detected during the untwist and check configuration stage (Fig. 2g-h).

Following a successful grasp of the second corner, a dynamic maneuver in which the towel is repeatedly pulled taut is used to bring the towel to a low-energy configuration that may still be partially twisted but can be corrected solely by in-place wrist roll motions (Fig. 2g-h). It does not appear feasible to avoid twisting the towel in the first place, as that would require accurately estimating the complete configuration of the towel while it is still hanging (and then constraining the arm trajectories to ones that leave the towel untwisted). A local search procedure is used to fully untwist both the left and right sides separately by maximizing an objective that is a linear combination of the gripper spread distance when the towel is pulled taut and the height to which the towel hangs on the side being corrected (estimated using the high-resolution stereo pair). The untwisting procedure also estimates the 3-D size of the towel (and also whether it is being held by a long side or short side) by fitting a 3-D rectangle to the foreground points; a sufficiently poor fit even in the optimally untwisted configuration indicates a mis-grasp, in which case the corner grasping is re-attempted.

For the actual folding, assuming it is held by a short side, the towel is held taut and pulled across the edge of the table in order to spread it out (Fig. 2k). The 3-D positions of the corners of the towel on the table are estimated based on a combination of foreground-background segmentation, stereo correspondence, and the (known) planar geometry of the table. This allows the robot to precisely align the two corners held in its grippers with the two corners on the table and then release, producing the first fold (Fig. 2l).[5] The towel is then re-grasped (Fig. 2m), folded a second time (Fig. 2n), and then finally re-grasped and placed on the stack (Fig. 2o). In the case that the towel is originally held by a long side, the table is used to spread out and regrasp the towel in the short side configuration, from which point folding proceeds as if the short side had been held originally.

An optimized GPU implementation of a variational algorithm for dense optical flow [20] was used by our implementation of the proposed grasp point detection algorithm for both optical flow computation as well as dense stereo matching.[6] As an optimization, the foreground-background segmentation was used to restrict optical flow computation to the foreground portion of the image, which was rescaled, if necessary, to not exceed 2 megapixels (to limit computation time). Because the dependency structure of the grasp point detection algorithm is not a simple path, it was executed more efficiently in a parallel pipelined fashion. A pipeline depth of 4 (meaning at most 4 frames were processed simultaneously) was used to maximally utilize the NVIDIA GTX 295 GPU used for the optical flow computation and the Intel Core 2 quad-core 2.5GHz CPU used for all other computation.

[5]Although the size of the towel has been estimated and is used to parameterize the folding (specifically the motions involved in pulling the towel across the table edge), due to the deformable nature of the towel, visual feedback has been necessary in order to achieve any reasonable result.

[6]Because this implementation does not compute a confidence measure along with the optical flow, a confidence measure is derived by computing the flow in both directions and defining the confidence $C(u,v) = e^{-0.2d(u,v)}$, where $d(u,v)$ is the round-trip displacement distance by following the flow forward and then backward at u,v.

Fig. 2. The state machine model of the procedure: dashed lines indicate failure recovery cases. The images show an actual run.

Background subtraction and mobile base localization were treated as solved problems, as they were not our research focus. Specifically, a known solid-color background was used to simplify foreground-background segmentation, and a motion-capture system was used for base localization.

V. EXPERIMENTS

After tuning the parameters on a training set of 3 towels[7], we ran 50 trials over a wide variety of previously unseen towels. In particular, the test set comprised 25 distinct colors/patterns, 16 distinct materials, and 11 distinct sizes (with lengths ranging from 32 to 78 cm and widths ranging from 32 to 48 cm).[8] Each of the 25 towels in the test set was fed to the robot individually by tossing it onto the table in a random configuration before beginning each test run. We tested each towel twice (each time in a different random configuration) for a total of 50 trials. Additionally, to verify handling of multiple towels at a time, the robot was tested on a pile of 5 towels (tossed onto the table in a random configuration).

We also used the data from the 50 trials to directly evaluate the performance of the grasp point detection algorithm and compare it with several baseline algorithms. The 1966 high-resolution stereo image pairs on which the algorithm was run over the course of these trials were saved (along with the output of the algorithm) and hand-labeled with the ground-truth locations of any corner grasp points.[9] (Only corners of the towel visible against the background in the image were labeled positive.) The final filtering based on stereo depth information and motion planning feasibility was excluded from this evaluation. The maximum-cardinality matching in each image between ground-truth corners and detected corners, with matches permitted only between points no more than 150 pixels apart (corresponding to about 3 cm), determined the true positive detections. Precision and recall were calculated based on the summed true positives, false positives, and false negatives in all images. To (approximately) reflect the low cost of obtaining an additional image from a different viewpoint and the high cost of recovering from a mis-grasp, the $F_{0.5}$-score was selected as a summary statistic for comparison purposes.[10]

The only methods that have previously been proposed specifically for cloth corner grasp point detection are based on selecting the bottommost point; by definition, such methods cannot, however, find any suitable grasp point for the second corner. Instead, therefore, a thresholded and non-maxima suppressed Harris detector [11] applied to the raw foreground-background edge map, and also the same detector applied instead to the thresholded border classification map (§III-B), were used as two baseline algorithms for evaluating the performance of the proposed grasp point detection algorithm.[11] The threshold parameter trades off recall for precision. The sensitivity parameter K and the size B of the square blocks used in computing the Harris response, along with the radius R of the circular region used for non-maxima suppression were selected individually for the two variants using a grid search to maximize the $F_{0.5}$-score (obtained from the optimal threshold) on a 386-image training subset (approximately 20%) of the labeled data that was set aside for this purpose. The training subset was not used for the proposed grasp point detection algorithm; it was used solely for the Harris detector-based algorithms. Performance comparisons were made using the testing subset of the labeled data, which comprised the remaining 1580 images.

VI. RESULTS

A. Grasp Point Detection Performance

The results of an evaluation using the dataset in its entirety, corresponding to the case of detecting the first corner in which any corner is acceptable, are shown in Fig. 3. Fig. 4 shows the results of additional evaluation corresponding to the detection of the second corner, namely excluding the bottommost corner from both detector output and the ground-truth labels and restricted to the subset of the data that was collected during attempts to detect the second corner. Given the greater importance of precision over recall, the proposed algorithm significantly outperforms the Harris detector baselines on both evaluations. As can be seen in the Harris detector results, the proposed border classification step substantially improves precision at the cost of only a modest reduction in recall. Furthermore, in combination with the border classification, the proposed RANSAC-based

[7]For the border classification algorithm, a dilation radius of $\delta = 40$ and a classification threshold of $t = 1.7$ for edges more than 35° from horizontal, and $t = 0$ otherwise, since horizontal folds are not stable on a free-hanging towel, were used. For the RANSAC corner detection, the parameters $d_{\text{inlier}} = 25$, $\ell_{\text{gap}} = 45$, and $\ell_{\text{min}} = 300$ were used. (Note that these parameter are relative to an image size of 3312x4416.) Grasps 5 cm diagonally in from the corner tip were used.

[8]Due to the limited range of motion of the arms and the limited field of view of the high-resolution cameras, the procedure is limited to towels no larger than about 78x48 cm; only towels within this range were included in the test set. Additionally, due to the use of a solid green pattern for background subtraction, green towels were also excluded from the test set.

[9]The data collection process, namely actual trial runs of the complete procedure, was influenced by the proposed grasp point detection algorithm. In particular, the fact that each corner detection attempt stops (and a grasp is attempted) as soon as a suitable detection is made negatively biases recall. This is mitigated, however, by the fact that many detections were excluded by a later stage of processing, e.g. due to motion planning infeasibility, and also due to the pipelining used by the detection algorithm, which typically caused 2 or 3 additional images to be captured after the image in which the first accepted detection was made. There is an additional bias towards lower recall due to the fact that towel configurations making detections most difficult were precisely the ones for which the most samples were collected; this was mitigated by reweighting the data to give equal total weight to each trial.

[10]The traditional F-measure (F_1-score) is the harmonic mean of precision and recall: $F_1 = 2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$. More generally we have $F_\beta = (1+\beta^2)\frac{\text{precision}\cdot\text{recall}}{\beta^2\cdot\text{precision}+\text{recall}}$. The $F_{0.5}$-score weights precision twice as much as recall.

[11]Actually grasping a corner using stereo to select a grasp position and orientation depends on a precise 2-D localization of the corner in the image, as produced by the proposed algorithm; the Harris detector produces only a point. An additional post-processing step would be needed to produce an estimate of the position and angle of the corner corresponding to the detected point. However, we evaluate the detector performance solely on the basis of point detections.

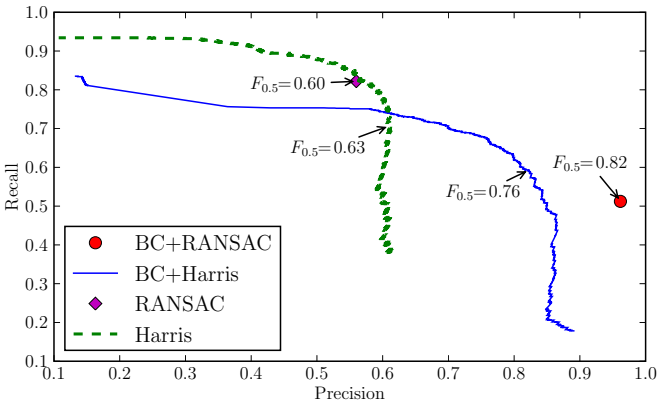

Fig. 3. Combined detection performance on all corners is compared using the testing subset. **BC+RANSAC** is the RANSAC-based corner detection(§III-C) applied to the border classification map(§III-B), namely the complete proposed algorithm without the final filtering based on stereo and motion planning feasibility(§III-D). **RANSAC** is the RANSAC-based corner detection applied directly to the foreground-background edge map, thus excluding the border classification step. **BC+Harris** is the Harris detector applied to the border classification map, using the optimized parameters $K = 0.2$, $B = 250$, $R = 800$. **Harris** is the Harris detector applied directly to the foreground-background edge map, using the optimized parameters $K = 0.2$, $B = 250$, $R = 600$.

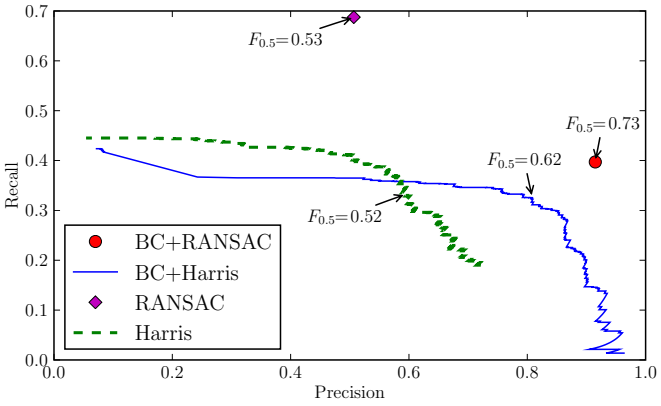

Fig. 4. Detection performance for the second corner only is compared using the testing subset. **BC+Harris** used the optimized parameters $K = 0.2$, $B = 300$, $R = 600$. **Harris** used the optimized parameters $K = 0.2$, $B = 300$, $R = 450$.

corner detection provides a substantial improvement over the Harris detector. The RANSAC-based corner detection does not outperform the Harris detector when not used in combination with the border classification, likely due to the fact that its parameters were tuned only for the combined use case, whereas the Harris detector was optimized separately on the training subset for each of the four cases.

B. Complete Procedure

A perfect success rate of 100% was achieved on the 50 end-to-end trials of previously untested towels. A total of 153 corner detection attempts were made over the course of the 50 trials, of which 147 led to a correct detection (there was one case of no detection) and 124 additionally led to a correct grasp. On a per-detection-attempt basis, this corresponds to a detection success rate of 96% and a combined detection

Step	Occur.	Dur./Occ.	Total Dur.
Initial Pickup	1.08	24	26
First grasp point	1.56	187	291
Second grasp point	1.50	348	521
Untwisting	1.30	186	242
Folding	1	397	397

Fig. 5. Running time summary of complete procedure on the 50 trials: average number of occurrences of each step per trial; average duration per occurrence of each step; and average total duration of each step per trial. The time spent actually grasping the detected grasp points, as well as the time spent on the shake maneuver, is included in the grasp point detection time, but only accounts for a small fraction of the time.

and grasping success rate of 81%. Although there were some failures of individual steps, as reflected by these modest but non-zero failure rates, the overall procedure was designed to be robust to such failures and indeed successfully recovered from all failures. The test on the pile of 5 towels was also completely successful. Videos of our autonomous folding runs are available at the URL provided in the introduction.

C. Running Time

A per-step breakdown of the running time of the complete procedure is shown in Fig. 5. The average total duration per towel (sum of the rightmost column) was 1478 seconds. The majority of time is spent on grasp point detection. The greater time spent on the second grasp point detection reflects the greater scarcity of suitable grasp points due to the exclusion of the bottommost corner. The speed of the pickup and folding steps was limited only by the speed at which the robot could follow the joint and base trajectories (which was not heavily optimized).

The overall throughput of the grasp detection algorithm was 1 frame per 13.8 seconds. For the first grasp point, an average of 6.9 frames were processed per attempt (before a grasp was attempted); for the second grasp point, an average of 16.9 frames were processed per attempt. Computationally, the throughput was primarily limited by the optical flow calculation: the NVIDIA GTX 295 GPU provided a throughput of 1 optical flow computation per 5 seconds, and the processing of each frame requires either one or two optical flow computations, depending on whether a 2-D grasp point detection is made.

D. Analysis of Failure Cases

On 28 of the 50 trials, no exceptional conditions occurred, meaning the robot correctly detected and grasped the first and second corners and then correctly untwisted and folded the towel without any intervening failures. On each of the remaining 22 trials, one or more of several types of recoverable failure conditions occurred:

Grasp attempts (on properly detected corners) that completely missed the towel were the most common type of failure, with 16 occurrences; they are also the least expensive in terms of added time, as they are detected immediately, allowing another detection attempt to be performed. These failures are primarily due to the inability of the blind grasping

strategy to account for the towel potentially being in a different configuration than when it was imaged (due to it hanging freely and possibly not settling for long enough or being in an unstable configuration); this limitation was exasperated by the use of pipelining to speed up the detection algorithm, as it results in additional rotations (and an opposite backward rotation) prior to grasping. (The stereo to arm calibration, accurate to about $1\mathrm{cm}$, was not a significant factor.) Overall, though, the blind grasping strategy proved to be quite adequate, failing due to lack of visual feedback on only 11% of grasps.

In 5 cases, the detection algorithm returned a false positive (typically due to a sharp fold in the cloth). Two of these cases happened to lead to a missed grasp; the other three cases led to successful grasps which were later correctly detected as mis-grasps during the untwisting and checking stage.

In 4 cases, a corner was correctly detected and grasped, but an additional nearby part of the towel happened to get caught in the gripper. These cases were later correctly detected as mis-grasps during the configuration checking stage.

In 3 cases, due to an insecure initial pickup grasp, the towel slipped out of the gripper (falling back onto the pile) while detecting the first corner (and then retried).

In 1 case, after correctly grasping the first corner, a suitable second corner grasp point was not detected, leading to the towel being dropped and retried.

In 3 cases, the untwisting stage failed to fully untwist the towel; these cases were subsequently correctly detected by the configuration checking stage.

In 5 cases, the configuration checking stage incorrectly classified a correctly grasped towel as being mis-grasped, leading to an unnecessary retry. These 5 false negatives out of a total 65 times that the configuration checking was invoked during the 50 trials correspond to a precision of 92%. Note that false positives by the configuration checking are not recoverable, but none occurred.

VII. Conclusions

We proposed a cloth grasp point detection algorithm which has been shown to have very high precision and a very reasonable rate of recall while being highly robust to variation in material, size, and appearance due to relying only on geometric cues. The reliability and robustness of our algorithm enabled for the first time a robot with general purpose manipulators to reliably and fully-autonomously fold previously unseen towels, demonstrating success on all 50 out of 50 single-towel trials as well as on a pile of 5 towels. Although our complete folding procedure was specialized to towels, the proposed algorithm could likely be useful for detecting grasp points on many types of clothing.

Acknowledgments

This work was supported in part by NSF under award IIS-0904672. We give warm thanks to our collaborators at Willow Garage for giving us the opportunity to work with their robotic platform and for their valuable input during our experiments.

References

[1] D.J. Balkcom and M.T. Mason. Introducing robotic origami folding. In *Proc. ICRA*, 2004.

[2] D. Berenson and S. Srinivasa. Grasp synthesis in cluttered environments for dexterous hands. In *IEEE-RAS International Conference on Humanoid Robots*, December 2008.

[3] D. Berenson, S. Srinivasa, D. Ferguson, A. Collet Romea, and J. Kuffner. Manipulation planning with workspace goal regions. In *Proc. ICRA*, May 2009.

[4] A. Bicchi and V. Kumar. Robotic grasping and contact: a review. In *Proc. ICRA*, 2000.

[5] D.L. Bowers and R. Lumia. Manipulation of unmodeled objects using intelligent grasping schemes. *IEEE Trans. on Fuzzy Systems*, 2003.

[6] J. Coelho, J. Piater, and R. Grupen. Developing haptic and visual perceptual categories for reaching and grasping with humanoid robot. *Robotics and Autonomous Systems*, 2001.

[7] N. Fahantidis, K. Paraschidis, V. Petridis, Z. Doulgeri, L. Petrou, and G. Hasapis. Robot handling of flat textile materials. *Robotics & Automation Magazine, IEEE*, 4(1):34–41, Mar 1997.

[8] 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, 1981.

[9] K. Gopalakrishnan and K. Goldberg. D-space and deform closure grasps of deformable parts. *The Int. J. of Robotics Research*, 24:899, 2005.

[10] K. Hamajima and M. Kakikura. Planning strategy for task of unfolding clothes. In *Proc. ICRA*, volume 32, pages 145–152, 2000.

[11] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[12] I. Kamon, T. Flash, and S. Edelman. Learning to grasp using visual information. In *Proc. ICRA*, 1996.

[13] C.C. Kemp, A. Edsinger, and E. Torres-Jarra. Challenges in robot manipulation in human environments. *IEEE Robotics and Automation Magazine*, 2007.

[14] Y. Kita, F. Saito, and N. Kita. A deformable model driven visual method for handling clothes. In *Proc. ICRA*, 2004.

[15] F. Lamiraux and L.E. Kavraki. Planning paths for elastic objects under manipulation constraints. *International Journal of Robotics Research*, 20(3):188–208, 2001.

[16] T. Matsuno and T. Fukuda. Manipulation of flexible rope using topological model based on sensor information. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 2638–2643, Oct. 2006.

[17] M. Moll and L.E. Kavraki. Path planning for deformable linear objects. *IEEE Trans. on Robotics*, 22(4):625–636, Aug. 2006.

[18] G.J. Monkman. Robot grippers for use with fibrous materials. *Int. J. Rob. Res.*, 14(2):144–151, 1995.

[19] A. Morales, P.J. Sanz, and A.P. del Pobil. Vision-based computation of three-finger grasps on unknown planar objects. In *Proc. IROS*, 2002.

[20] M.Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. British Machine Vision Conference (BMVC)*, London, UK, September 2009.

[21] F. Osawa, H. Seki, and Y. Kamiya. Clothes folding task by tool-using robot. *Journal of Robotics and Mechatronics*, 2006.

[22] F. Osawa, H. Seki, and Y. Kamiya. Unfolding of massive laundry and classification types by dual manipulator. *JACIII*, 11(5):457–463, 2007.

[23] R. Platt, R.A. Grupen, and A.H. Fagg. Re-using schematic grasping policies. In *Humanoids*, 2005.

[24] R. Platt, R.A. Grupen, and A.H. Fagg. Learning grasp context distinctions that generalize. In *Humanoids*, 2006.

[25] K. Salleh, H. Seki, Y. Kamiya, and M. Hikizu. Inchworm robot grippers in clothes manipulation optimizing the tracing algorithm. In *Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on*, pages 1051–1055, Nov. 2007.

[26] A. Saxena, J. Driemeyer, and A. Ng. Robotic grasping of novel objects using vision. *International Journal on Robotics Research*, 2008.

[27] K.B. Shimoga. Robot grasp synthesis algorithms: A survey. *International Journal on Robotics Research*, 1996.

[28] K. Wyrobek, E. Berger, H.F.M. Van der Loos, and K. Salisbury. Towards a personal robotics development platform: Rationale and design of an intrinsically safe personal robot. In *Proc. ICRA*, 2008.

[29] K. Yamakazi and M. Inaba. A cloth detection method based on image wrinkle feature for daily assistive robots. In *IAPR Conf. on Machine Vision Applications*, pages 366–369, 2009.