Active Perception of Deformable Objects using 3D Cameras

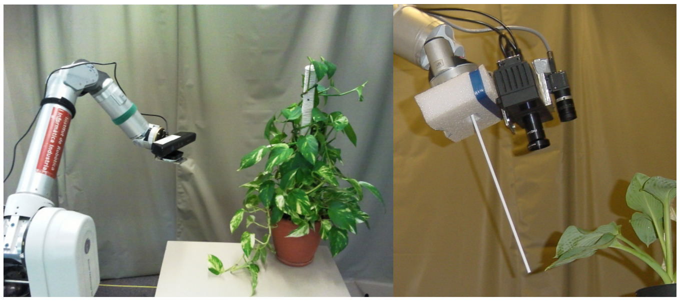
Guillem Alenyà, Francesc Moreno-Noguer, Arnau Ramisa, and Carme Torras

Abstract— **Perception and manipulation of rigid objects has received a lot of attention, and several solutions have been proposed. In contrast, dealing with deformable objects is a relatively new and challenging task because they are more complex to model, their state is difficult to determine, and self-occlusions are common and hard to estimate. In this paper we present our progress/results in the perception of deformable objects both using conventional RGB cameras and active sensing strategies by means of depth cameras. We provide insights in two different areas of application: grasping of textiles and plant leaf modelling.**

I. INTRODUCTION

3D perception of deformable objects using RGB cameras has been one the most studied research fields within computer vision. There exist a large number of techniques for this purpose, such as stereovision, shape from shading, structure-from-motion or shape from texture. For robotics applications, monocular techniques that require one acquisition are probably the most interesting approaches, because they avoid the occlusion problems that appear when dealing with multiple views and a single camera may be easily incorporated on a robotic arm, for instance. On the negative side, retrieving non-rigid shape using one single image is a highly ambiguous problem, because many different shapes may have similar projections. In our group, we have researched on techniques for addressing this [1], [2].

On the other hand, the problem may be highly simplified when using the now popularized 3D cameras. The technology of 3D cameras has quickly evolved in recent years, yielding off-the-shelf devices with great potential in many scientific fields ranging from virtual reality to surveillance and security. In particular within robotics, these cameras open up the possibility of real-time robot interaction in human environments, by offering an alternative to computationally costly procedures such as stereovision and laser scanning. Time-of-Flight (ToF) cameras, provided by Mesa Imaging and PMD Technologies among others, appeared first and attracted a lot of attention with dedicated workshops (e.g., within CVPR'08) and a quickly growing number of papers at major conferences. These days the appearance among others of the Kinect camera, with the Light Coding technology provided by PrimeSense and based on Structured Light (SL), has received even greater attention, because of its low cost and simplicity of use.



(a) Kinect camera on the robot

(b) Detail of the ToF+color camera sensor

Fig. 1: Experimental setup with the robot arm used in the experiments in two different configurations.

We have studied the use of ToF cameras to assist robot learning of manipulation skills in a kitchen environment. Since this entailed mobile manipulation of rigid objects guided by a human teacher, we surveyed near one hundred previous works in three scenarios of application, namely scene-related tasks involving mobile robots in large environments, object-related tasks entailing robot interaction at short distances, and human-related tasks dealing with face, hand and body recognition for robot-human interfaces [3]. Our conclusion was that ToF cameras seem especially adequate for mobile robotics and real-time applications in general, and in particular for the automatic acquisition of 3D models requiring sensor motion and on-line involved computations, which was the target application finally developed [4].

We now have interest in two different scenarios involving deformable objects. One is the perception of textiles to estimate adequate grasping points. In the context of the PAU project [5] perception and manipulation of deformable objects is investigated, as the problem is challenging considering its high dimensionality and the difficulties related to the uncertainty.

The other scenario is aimed at enhancing the perception of plants. The ongoing project GARNICS [6] aims at automatically monitoring large botanic experiments so as to determine the best treatments (watering, nutrients, sunlight) to optimize predefined aspects (growth, seedling, flowers) and eventually guiding robots, like the one in Figure 1, to interact with plants in order to obtain samples from leaves to be analyzed or even to perform some pruning. Here the interest is focused on 3D model acquisition of deformable objects (leaves) and their subsequent manipulation.

Color vision is helpful to extract some relevant plant features, but it is not well-suited for providing the struc-

This research is partially funded by the EU GARNICS project FP7-247947, by the Spanish Ministry of Science and Innovation under project DPI2008-06022 and MIPRCV Consolider Ingenio CSD2007-00018, and the Catalan Research Commission.

Authors are with Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain; {galenya,fmoreno,aramisa,torras}@iri.upc.edu

tural/geometric information indispensable for robot interaction with plants. 3D cameras are, thus, a good complement, since they directly provide depth images. Moreover, plant data acquired from a given viewpoint are often partial or ambiguous, thus planning the best next viewpoint becomes an important requirement. This, together with the need of a high throughput imposed by the application, makes 3D cameras (which provide images at more than 25 frames-per-second) a good option in front of other depth measuring procedures, such as stereovision or laser scanners. Since now ready-to-use SL cameras are also available, we undertook a comparative assessment of the usefulness of both ToF and SL cameras to acquire (possibly deformable) object models at close distances and to calibrate them with respect to the robot for subsequent manipulation.

The paper is structured as follows. First we present our advances in reconstructing deformable objects using one single RGB camera. Then, in Sec. III we present two different depth camera technologies: ToF and SL. The first area of application, grasping of textiles, is described in Sec. IV. Active vision, with the camera mounted on a robotic arm, is presented in Sec. V in relation to the botanic application. Finally, Sec. VI is devoted to some discussions about the results and possible exploitation of these technologies.

II. Non-Rigid Reconstruction using a Single RGB Camera

It has been shown that the 3D shape of deformable surfaces can be very effectively recovered from even single images provided that enough correspondences can be established between that image and one in which the surface's shape is already known [7], [1], [8]. While effective, these techniques only return one reconstruction without accounting for the fact that several plausible shapes could produce virtually the same projection and therefore be indistinguishable on the basis of correspondences and geometry alone. In practice, as shown in Fig. 2, disambiguation is only possible using additional information, such as that provided by shading patterns.

In [2], we introduced an effective way to sample the space of all plausible solutions. We achieved this by representing shape deformations in terms of a weighted sum of deformation modes and relating uncertainties in the location of point correspondences to uncertainties in the mode weights. This let us explore the space of modes and, in the end, select a very small number of likely ones, which correspond to 3D shapes such as those depicted in the top row of Fig. 2.

In practice, to select the best one, we used lighting information that comes from either distant or nearby light sources. The latter was particularly significant because exploiting it would involve solving a difficult non linear minimization problem if we did not have a reliable way to generate 3D shape hypotheses. In our examples, this was all the more true since the lighting parameters are initially unknown and had to be estimated from the images. This also means that we could have used other sources of shape information besides shading. We showed that these approaches outperformed state-of-the-art methods [9], [1].A few sample frames of the results are shown in Fig. 3.

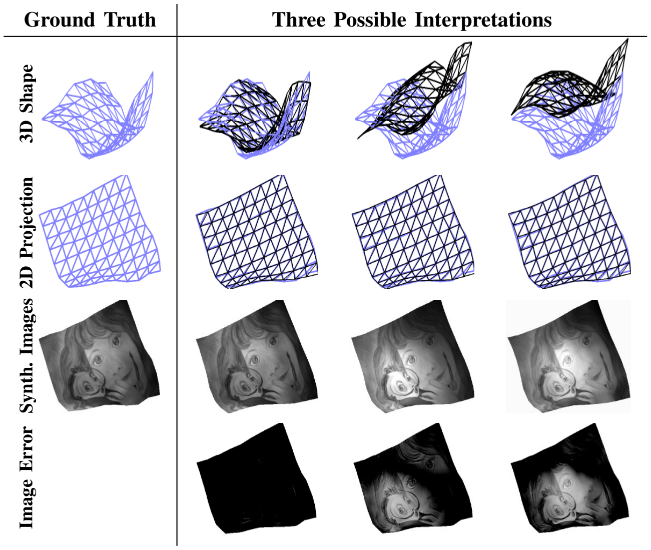


Fig. 2: Handling 3D shape ambiguities. **Left Column.** An image of a surface lit by a nearby light source and the corresponding ground truth surface. **Three other Columns.** In each one, a different candidate surface proposed by our algorithm is shown in black. The corresponding projection and synthesized image given automatically estimated lighting parameters are shown below. As can be seen in the second row, its projection is very similar, even though its shape may be very different from the original one. In other words, the candidates cannot be distinguished based on reprojection error alone. However, when comparing the true and synthesized images, it becomes clear that the correct shape is the one at the top of the second column.

In other words, our contribution was an approach to avoiding being trapped in the local minima of a potentially complicated objective function by efficiently exploring the solution space of a simpler one. As a result, we only had to evaluate the full objective function for a few selected shapes, which implied we could use a very discriminating one if necessary.

In the following sections we will turn to other approaches that instead of capturing the 3D structure using RGB cameras, directly use the information of depth sensors. Although RGB cameras offer a more general solution that may potentially be used in unconstrained and outdoor environments, the depth sensors represent a robust solution specially in situations where lighting may be controlled.

III. Depth cameras

We will consider two different 3D camera types, a CamCube ToF camera and the Kinect sensor.

ToF camera is a relatively new type of sensor that delivers 3-dimensional images at high frame rate, simultaneously providing intensity data and range information for every pixel. Figure 4 shows the depth image of a plant leaf with the depth values coded as different color values.

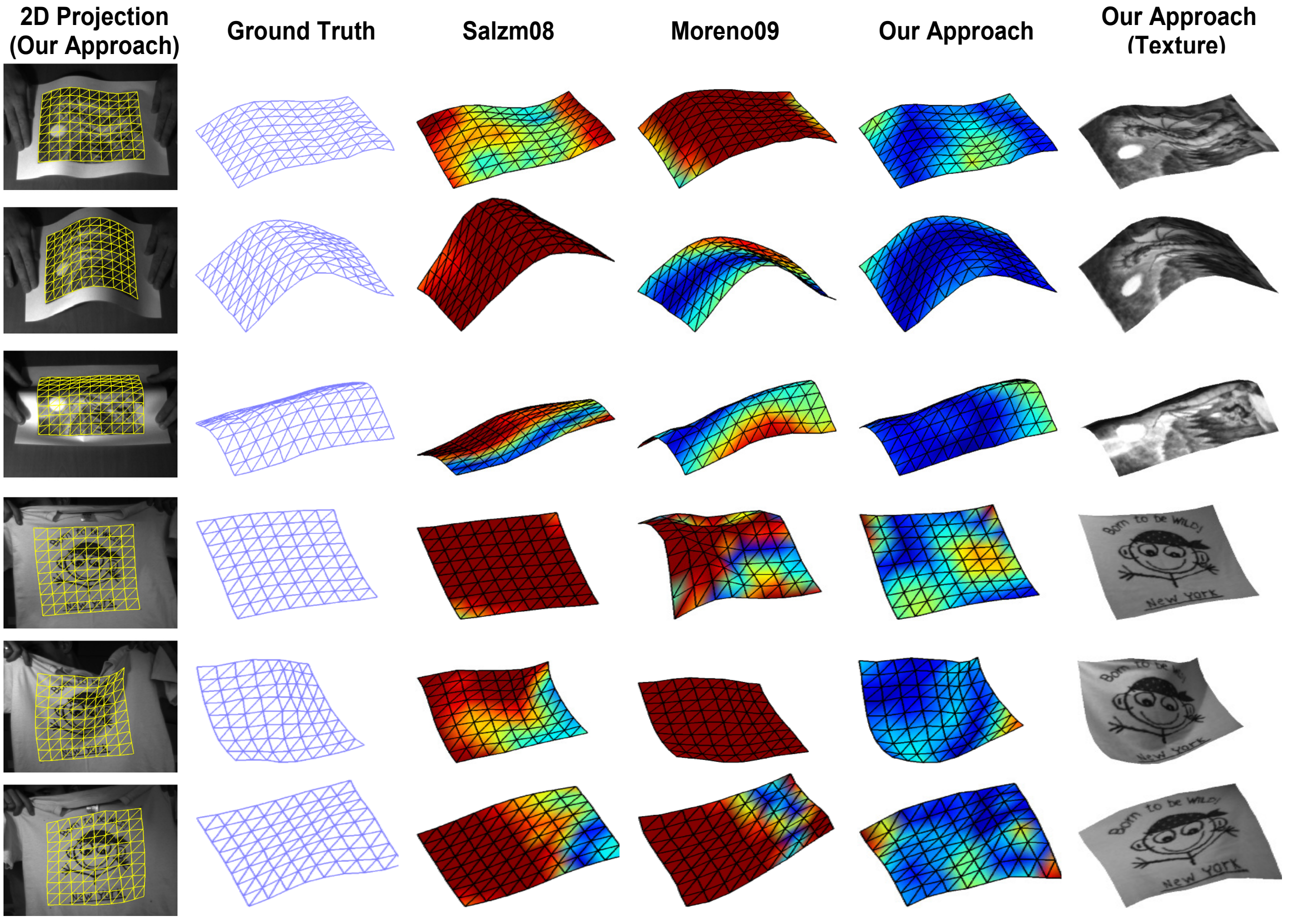


Fig. 3: Results of reconstructing shape from single RGB images using [2] and two other approaches [9], [1]. Top three rows: Results of a bending paper. Bottom three rows: Results of a deforming T-Shirt. Note that our results are consistently more accurate. The reconstruction figures are color coded, such that reddish areas represent regions with larger errors.

Depth measurements are based on the well-known time-of-flight principle [10]. A radio frequency modulated light field is emitted by the system and then reflected back to the sensor, which allows for the parallel measurement of its phase (cross-correlation), offset and amplitude [11].

Kinect uses an infrared structured light emitter to project a pattern into the scene and a camera to acquire the image of the pattern, then depth is computed by means of structured light algorithms. Additionally, among others sensors, the Kinect integrates a high resolution color camera.

Kinect was developed with the aim of robust interactive human body tracking and great efforts have been made in this direction [12]. After the Kinect protocol was hacked, the community rapidly started to use it, first with the same aim of human interaction and afterwards in other areas, like robot navigation[1]. Later, the official library was made public through the OpenNi organization.

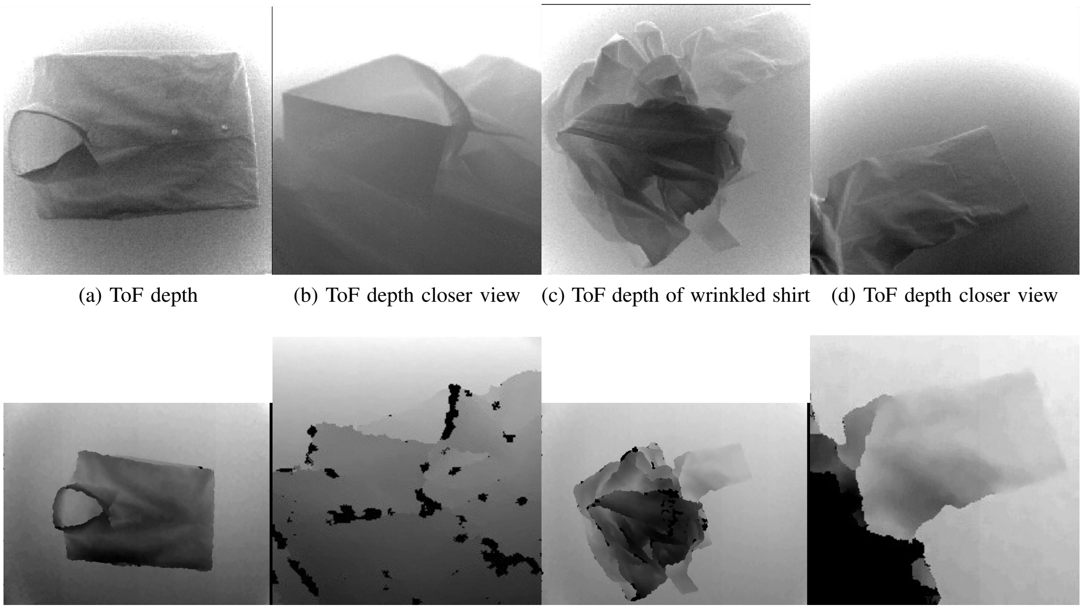
Both camera types can deliver depth images at reasonably high frame rates. The main difference is in resolution: ToF cameras still have limited resolution (typically around 200 x 200), while the Kinect depth camera exhibits VGA resolution. Both camera types are auto-illuminated so in principle they can work in a wide variety of illumination conditions.

One common problem with both cameras is that they do not provide a dense depth map. The delivered depth images contain holes corresponding to the zones where the sensors have problems, whether due to the material of the objects (reflection, transparency, light absorption) or their position (out of range, occlusions). As will be presented in the next sections, Kinect is more sensitive to this problem by construction.

To compare both cameras in one of our scenarios, we take several images of a shirt (Fig. 5) in different configurations. Both cameras offer good depth estimation of the shirt, and even small wrinkles can be identified. The close views with ToF and Kinect provide lots of details. Observe clearly the shape of the collar (Figs. 5b and 5f), the different depths in the top image of the wrinkled shirt (Figs. 5c and 5g), and the details of the shirt sleeve (Figs. 5d and 5h).

As regards to Kinect, in Figure 5f occlusions appear in the collar and this produces holes in the surface, presumably due to bad readings as no occlusions are present. We should

[1] See for example the initiative of commercially releasing a low-cost robot based on iRobot Create and Kinect at http://www.willowgarage.com/turtlebot


(a) ToF depth (b) ToF depth closer view (c) ToF depth of wrinkled shirt (d) ToF depth closer view

(e) Kinect depth (f) Kinect depth detail (g) Kinect depth of wrinkled shirt (h) Kinect depth detail

Fig. 5: Images of a folded and a wrinkled shirt. Images are obtained by moving both ToF and Kinect cameras to obtain the best possible depth acquisition. The wrinkles in the shirt, even if they are small, are visible with both cameras. (f) Observe the holes in some parts of the surface and the occlusions in the collar.

note that the position, size and number of holes vary with the sensor motion.

IV. Grasping Cloth Using Depth Information

Recently the problem of grasping and folding clothes with a robotic arm has attracted much attention [13], [14], [15], [16], [17], [18]. Its application ranges from automatizing industrial cleaning facilities to domestic service robots.

There exist works devoted to determining the best/optimal grasping point for a particular purpose (e.g. folding) once the cloth is held by a robotic hand. However, most of the research done in this area has been used in controlled environments and simple heuristics have sufficed.

A common heuristic or workaround used by works addressing textile grasping, such as [15], [18], is to select as grasping point the highest one in the 3D point cloud of the cloth object. However, in practice, the highest point does not need to constitute a good grasping point for robotic manipulators.

We have investigated what constitutes a good initial grasping point for a piece of cloth lying on a flat surface in an arbitrary configuration. Below we propose a new "wrinkledness" measure [19] that uses range information that can be used to determine the most easily graspable point at an affordable computational cost. Compared to other works [14], we directly use 3D information obtained from a low-cost sensor, therefore avoiding the expensive data collection and manual annotation step required for SVM training, and which not being vulnerable to learning errors.

Our initial assumption is that a good grasping point for a textile object lying on a table is one where the cloth defines ridges or other 3D structures, i.e. there are wrinkles. The justification of this assumption comes from the nature of the grasping mechanism, which in our case has three fingers, with a total of four degrees of freedom. Lacking the precision of movement, flexibility and the small(er) size of human hands (which can pick up cloth objects from the edges), the best point for a grasp for this type of hand is a pyramidal or conic-like shape, such as the one produced by wrinkles.

We have developed a measure of the "wrinkledness" in a point taking into account the depth information of its neighbourhood. This measure is computed using a local descriptor based in the surface normals of a 3D point cloud. In particular, we use the *inclination* and *azimuth* angles defined in the spherical coordinates representation of the normal vectors:

$$(\phi, \theta) = \left(arccos\left(\frac{z}{r}\right), arctan\left(\frac{y}{x}\right)\right) \quad (1)$$

where ϕ is the inclination and θ is the azimuth, (x, y, z) are the 3D point coordinates, and r is the radius in spherical ones, defined as:

$$r = \sqrt{x^2 + y^2 + z^2} \quad (2)$$

Next, we model the distribution of the inclination and azimuth values in a local region around each point. A beneficial side effect of this process is that occluded regions and areas where the Kinect was not able to estimate the depth are naturally interpolated using the information provided by

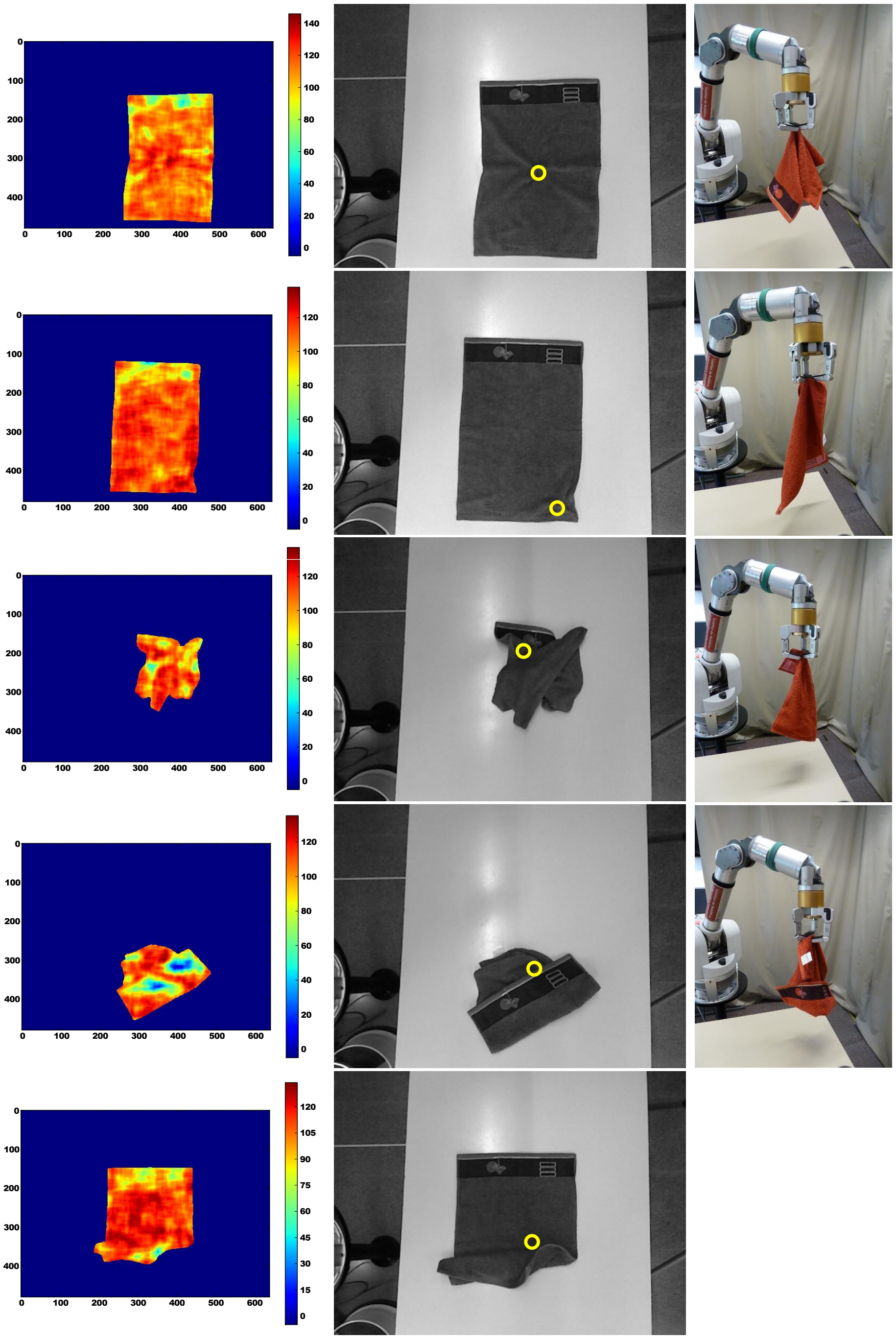
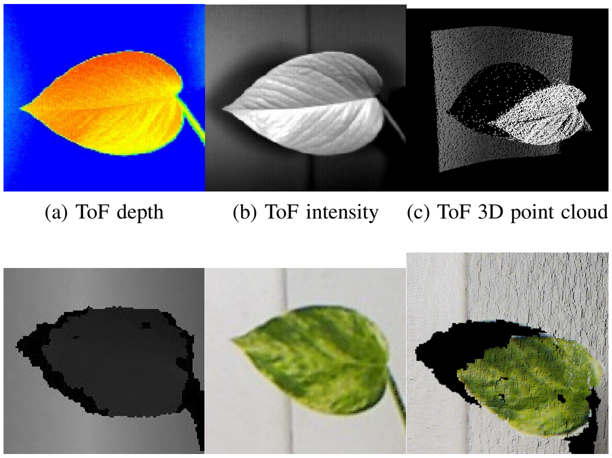


Fig. 6: Details for the five experiments conducted with a robotic arm (one per row). For each experiment is shown (in order): the segmented "wrinkledness" map of the towel, the selected grasping point, and a picture of the robotic hand with the grasped towel, if successful.

their neighbours, which reduces the sparsity of the point cloud.

From this model of the local distribution of normal angles in spherical coordinates, we seek to estimate the "wrinkledness" of a point. This can be intuitively done by looking at the spread of the angle histogram: the more different orientations the surface takes, the more likely that it is a highly wrinkled area. Although standard deviation is probably the first measure of spread that comes to mind, it is not a good choice, since a strongly bimodal distribution can have a large standard deviation while having low spread. A better choice is entropy, which does not suffer from this


(a) ToF depth (b) ToF intensity (c) ToF 3D point cloud

(d) Kinect depth (e) Kinect color (f) Kinect 3D point cloud

Fig. 4: Typical images supplied by a ToF camera and a Kinect camera. Figures (c) and (f) are the reconstructed 3D point clouds for each camera. (c) Observe the false *flying points* points between the leaf edge and background. (d) Observe the holes between the leaf and the background due to occlusions between the IR projector and the camera.

drawback:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i)\,, \qquad (3)$$

where X is the n-bin angle orientation histogram, and x_i is the i_{th} bin.

We tested our proposed "wrinkledness" measure in real grasping experiments. Our experimental setup consists of a robotic hand with three fingers installed in front of a flat table of uniform color, in which a small red towel was randomly positioned.

In all the experiments, a 2D histogram with a square support region with a side of 33 pixels was used to generate the "wrinkledness" map after segmenting the towel from the table, and the point with the highest activation was selected as the grasping point. Next the robotic arm was moved to the point, and a grasp attempt was performed. Please note that we are not claiming that the point with highest activation in the map is necessarily the best possible grasping point. However, we have used this simple heuristic with very good results.

Four out of five tests ended with a successful grasp. Figure 6 shows the images and "wrinkledness" maps used to decide the grasping point, and a photo of the robotic arm holding the towel for those tests that were successful. In each successive test the towel was positioned in increasingly difficult configurations.

V. Next best view and tracking

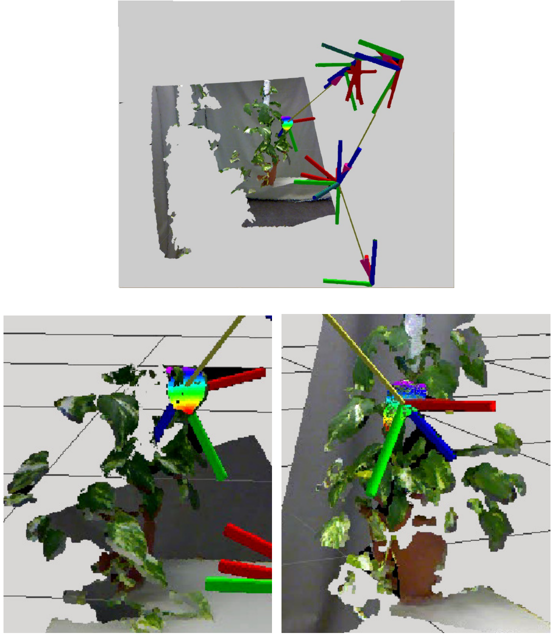


Fig. 7: Frames of a leaf tracking experiment. The set of connected reference systems represent the current position of the robot. The 3D points of the tracked leaf are colored with the depth and an additional reference system is attached to the leaf with Z coordinate (blue) normal to the leaf surface.

Recently we have presented a work on next view selection for plants [20]. The algorithm first selects some candidate plant leaves from a initial image, then extracts some geometrical characteristics and use them to move a combined ToF+color camera sensor with a robotic arm to obtain new and more detailed views of the selected leaves. Our approach uses a combination of depth and color information to perform image segmentation and robot guidance, and use some characteristics of the point cloud to extract the contours to segment the depth image [21].

We are now interested not only in the first general image and the last detailed image, but also in the sequence of images acquired while the robot is moving. This allows to perform a guided segmentation of the leaf in the final position, as well as continuously updating the 3D model of the plant using an uncertainty reduction algorithm [4].

In contrast to our previous work, we present here some results using a Kinect camera (Fig. 7). The experimental setup is shown in Figure 1a, where the camera is mounted on the end-effector of a WAM robot arm. Here we show the tracking using 3D information, so a leaf is manually selected and the robot arm is moved trying to keep the leaf into the image area. The real-time tracking uses a geometrical model of the leaf [22] and the central position and the normal orientation are extracted. In Fig. 7 the leaf points are colored depending on the depth, and the reference system is attached to the computed leaf central point with the Z component (in blue) normal to the surface at the center point[2].

As explained before, using a Kinect camera it is not

[2] The complete video can be accessed at http://www.iri.upc.edu/people/galenya/pub/LeafTracking.avi

possible to approach the leaf in the same manner as we did with the ToF+color combination, this being the reason why approaching motions are quite restricted.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have presented some recent work of our group towards perception and manipulation of deformable objects.

We are convinced that in these scenarios having depth information of the scene is crucial to produce robust and repetitive algorithms. Apart from the classical stereo and range finders sensors, we have extensive experience with the two different camera technologies that we have presented: ToF cameras and SL cameras. Although ToF cameras have lower resolution, they can provide depth images at short distances of up to 20cm. This capability makes them very valuable in contexts where fine details on the objects are crucial.

Two different areas of application have been presented. The first one is manipulation of textiles. We have presented some preliminary work towards finding a good measure of "graspability" for cloth objects lying on a flat surface. This is an important aspect for making robots fully autonomous in unprepared environments; in contrast, related literature so far relied on simple heuristics that worked in controlled settings.

One important limitation of this approach is that concave areas of the image get a high activation level while not being good grasping points. Yet, it is possible to compute a concavity measure and use it to re-weight the "wrinkledness" map.

As next step, we think that better grasping points could be found by combining information like point height, total 3D volume, normal orientation or the aforementioned concavity measure with the entropy-based measure proposed in this paper.

The second area of application is plant monitoring. Food industry is very important for society, and current efforts in automation are devoted to monitoring and performing actions on individual plants belonging to large plantations. Our leaf tracking example has been developed using a Kinect camera, yielding a very robust performance under varying conditions, since the precision requirements were relatively low. On the contrary, in the past we have also used a ToF camera under a next-best-view approach to find suitable leaves from which to take probes. Since this requires getting very close to the plant and finding suitable probing points with high precision, a ToF camera was more appropriate, although it required considerable parameter tuning.

Plants evolve with time, change their shape and their topology. We are exploring now how to create complete models of a plant, and how these models should be updated with time. An important aspect of the modeling process is to create models containing enough information to allow robotized interaction with the plant, for example cutting some leaves or taking probes for posterior analysis.

REFERENCES

[1] F. Moreno-Noguer, M. Salzmann, V. Lepetit, and P. Fua, "Capturing 3D stretchable surfaces from single images in closed form," in *Proc. 23rd IEEE Conf. Comput. Vision Pattern Recog.*, Florida, Jun. 2009, pp. 1842–1849.

[2] F. Moreno-Noguer, J. Porta, and P. Fua, "Exploring ambiguities for monocular non-rigid shape estimation," in *Proc. 11th European Conf. Comput. Vision*, 2010, pp. 361–374.

[3] S. Foix, G. Alenyà, and C. Torras, "Exploitation of Time-of-Flight (ToF) cameras," IRI, UPC, Tech. Rep. IRI-DT-10-07, 2010.

[4] S. Foix, G. Alenyà, J. Andrade-Cetto, and C. Torras, "Object modeling using a ToF camera under an uncertainty reduction approach," in *Proc. IEEE Int. Conf. Robot. Automat.*, Anchorage, May 2010, pp. 1306–1312.

[5] "PAU: Perception and action under uncertainty," 2008. [Online]. Available: http://www.iri.upc.edu/research/webprojects/pau/

[6] "GARNICS: Gardening with a cognitive system," 2010. [Online]. Available: http://www.garnics.eu/

[7] M. Perriollat, R. Hartley, and A. Bartoli, "Monocular template-based reconstruction of inextensible surfaces," in *Proc. British Machine Vision Conf.*, 2008.

[8] J. Zhu, S. Hoi, Z. Xu, and M. Lyu, "An effective approach to 3D deformable surface tracking," in *Proc. 10th European Conf. Comput. Vision*, ser. Lect. Notes Comput. Sci., vol. 5302, Marseille, 2008, pp. 766–779.

[9] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua, "Closed-form solution to non-rigid 3D surface registration," in *Proc. 10th European Conf. Comput. Vision*, ser. Lect. Notes Comput. Sci., vol. 4, Marseille, 2008, pp. 581–594.

[10] G. A. S. Foix and and C. Torras, "Lock-in time-of-flight (tof) cameras: a survey," *IEEE Sensors J.*, 2011.

[11] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *IEEE J. Quantum Electron.*, vol. 37, no. 3, pp. 390–397, Mar. 2001.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from a single depth image," in *Proc. 25th IEEE Conf. Comput. Vision Pattern Recog.*, Colorado Springs, Jun. 2011, to Appear.

[13] F. Osawa, H. Seki, and Y. Kamiya, "Unfolding of massive laundry and classification types by dual manipulator," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 11, no. 5, pp. 457–463, 2007.

[14] 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*, 2009, pp. 366–369.

[15] 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 *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2308–2315.

[16] M. Cusumano-towner, A. Singh, S. Miller, J. F. O. Brien, and P. Abbeel, "Bringing Clothing into Desired Configurations with Limited Perception," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, Shangai, China, 2011, pp. 3893–3900.

[17] S. Miller, M. Fritz, T. Darrell, and P. Abbeel, "Parametrized Shape Models for Clothing," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*, Shangai, China, 2011, pp. 4861–4868.

[18] B. Willimon, S. Birchfield, and I. Walker, "Classification of Clothing using Interactive Perception," in *Proc. IEEE International Conference on Robotics and Automation (ICRA11)*, 2011, pp. 1862–868.

[19] A. Ramisa, G. Alenyà, F. Moreno-Noguer, and C. Torras, "Determining where to grasp cloth using depth information," in *Proc. 14th Int. Conf. Catalan Assoc. Artificial Intell.*, Lleida, Oct. 2011.

[20] G. Alenyà, B. Dellen, and C. Torras, "3d modelling of leaves from color and tof data for robotized plant measuring," in *Proc. IEEE Int. Conf. Robot. Automat.*, Shanghai, May 2011, pp. 3408–3414.

[21] S. Foix, G. Alenyà, and C. Torras, "Towards plant monitoring through next best view," in *Proc. 14th Int. Conf. Catalan Assoc. Artificial Intell.*, Lleida, Oct. 2011.

[22] B. Dellen, G. Alenyà, S. Foix, and C. Torras, "Segmenting color images into surface patches by exploiting sparse depth data," in *Winter Vision Meeting: Workshop on Applications of Computer Vision*, Kona, Hawaii, 2011, pp. 591–598.