# FINDDD: A Fast 3D Descriptor to Characterize Textiles for Robot Manipulation

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

***Abstract*— Most current depth sensors provide 2.5D range images in which depth values are assigned to a rectangular 2D array. In this paper we take advantage of this structured information to build an efficient shape descriptor which is about two orders of magnitude faster than competing approaches, while showing similar performance in several tasks involving deformable object recognition. Given a 2D patch surrounding a point and its associated depth values, we build the descriptor for that point, based on the cumulative distances between their normals and a discrete set of normal directions. This processing is made very efficient using integral images, even allowing to compute descriptors for every range image pixel in a few seconds. The discriminative power of our descriptor, dubbed FINDDD, is evaluated in three different scenarios: recognition of specific cloth wrinkles, instance recognition from geometry alone, and detection of reliable and informed grasping points.**

## I. Introduction

Manipulation of highly deformable textile objects in unconstrained domestic environments is an ability with which robotic assistants for our homes should be equipped, so they could perform tasks such as laundry handling, ironing, or garment folding. Yet, this goal is not easy: the flexibility of textiles and the ambiguity of their appearance makes it very difficult to estimate their state at a given time, and therefore to plan the appropriate actions to carry out the selected task.

Despite its intrinsic difficulty, recent years have seen significant progress in this field. Some works, like Martin-Shepard et al. [1] and Cusumano-Towner et al. [2], demonstrated functional end-to-end systems (albeit in very controlled settings). The proposed systems are able to pick up a piece of laundry and manipulate it until a desired configuration is reached. Other approaches are more focused on the perception capabilities: Miller et al. [3] proposed a method based on parametrized shape models for estimating the pose of a crudely spread cloth item. Ramisa et al. [4] combined appearance and shape descriptors in a bag of features framework for the task of detecting collars of polo shirts. Willimon et al. [5] exploited the manipulation capabilities of a robot to help in a perception task, and proposed a system to pick up the topmost element of a pile of clothing, and subsequently classified it using interactive perception and four basic visual features. However, most of these works suffer from two important drawbacks that limit their applicability in everyday environments: a high computational cost, and perception methods usually designed *ad hoc* for particular experimental setups.

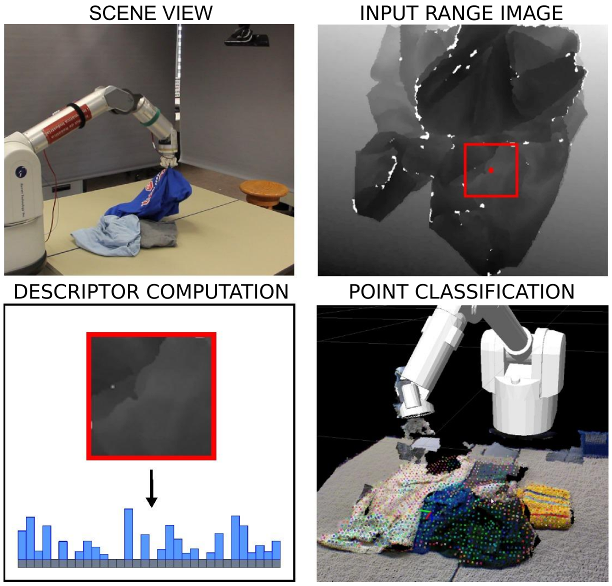


Fig. 1: Robotic textile manipulation requires efficient perception systems to account for the quick configuration changes a cloth can undergo. We propose the FINDDD shape descriptor for 2.5D range images, that can be computed very fast. Its effectiveness is demonstrated in different robotic tasks, such as that of recognizing good grasping points on a polo shirt, as shown in the figure.

In this work we address these two challenging limitations by proposing the Fast Integral Normal 3D (FINDDD) shape descriptor, that is both discriminative and very fast to compute. This descriptor can be used as a basic building block for 3D perception algorithms in robotic tasks that require very quick decision processes. Textile manipulation is one of such domains, as the configuration of the garments can rapidly and arbitrarily change under the action of a robotic hand (see Figure 1 for some example images).

The core idea behind our approach is to use integral images to exploit the fact that current depth sensors provide 2.5D range measurements, i.e, depth values arranged in a structured 2D array. Then, given one such input range image, FINDDD describes each image patch based on the cumulative distances between the normal vectors in the patch, and a discrete set of normal directions that uniformly sample the upper hemisphere. Both the normal vectors and

This research is partially funded by the Spanish Ministry of Science and Innovation under project DPI2011-27510, by the EU Project IntellAct FP7-ICT2009-6-269959 and by the ERA-Net CHISTERA project ViSen PCIN-2013-047. A. Ramisa worked under CSIC/FSE JAE-Doc grant.

The authors are with the Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Llorens i Artigas 4-6, 08028 Barcelona, Spain (aramisa, galenya, fmoreno, torras) at iri.upc.edu

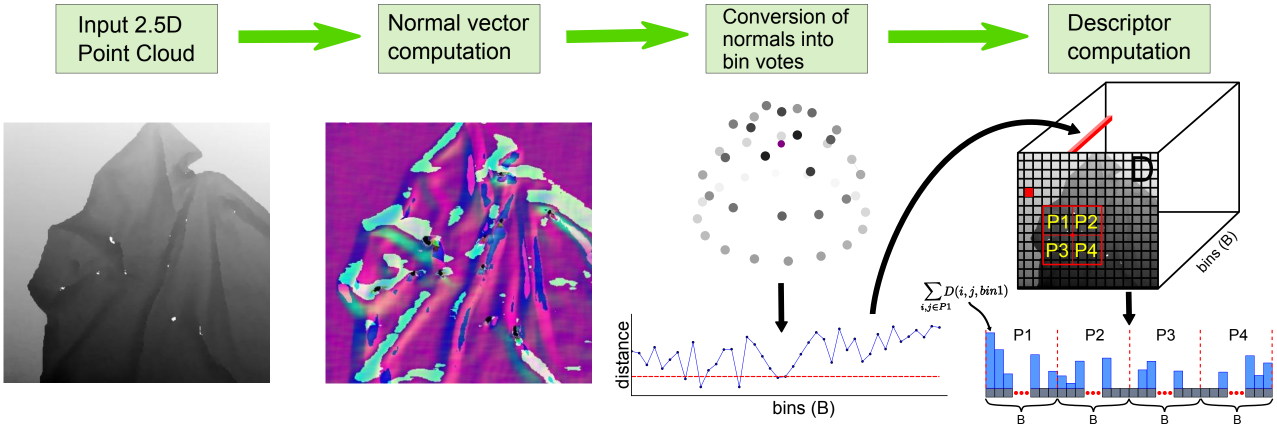

Input 2.5D Point Cloud
Normal vector computation
Conversion of normals into bin votes
Descriptor computation
D
P1 P2
P3 P4
bins (B)
$\sum_{i,j \in P1} D(i,j,bin1)$
distance
bins (B)
P1
P2
P3
P4
B
B
B
B

Fig. 2: Steps of the proposed method. The first two images show and input range image (darker means closer) and the normal vectors (mapped into RGB values) computed from that input image. The third image shows a 3D view of the orientation bin centers distributed on a unit hemisphere, with the input normal plotted as a purple dot. The gray level of the bin centers represents the distance between the normal and the center, which is linearly represented in the plot below. The red line in this plot indicates the distance above whose contribution to the bin centers is 0. The last image shows how the final descriptor is computed aggregating the votes for all the pixels within each sub-region.

the cumulative distances can be computed very efficiently with the use of integral images. To the best of our knowledge, this strategy has not been previously exploited in any of the recent descriptors typically used in robotics, such as NARF [6], SHOT [7], PFH [8], or FPFH [9]. On the negative side, when performing summations over rectangular domains we lose the 3D proximity constraints, as we might be merging the information of neighboring pixels at very different depths. Nevertheless, this is not critical in the domain of textile manipulation, as usually clothes are laying on top of a table and there are no strong depth discontinuities. In fact, as we will show in the experimental section, FINDDD yields recognitions rates that are very similar to those of the previously mentioned state-of-the-art descriptors, while being about two orders of magnitude faster.

Note that our claims with respect to the FINDDD descriptor are restricted to the problem of 3D perception for robotic textile manipulation, and that it has trade-offs which could degrade its performance in other domains such as multi-view rigid object detection or scene categorization.

## II. Previous 3D Descriptors

With the recent popularity of 3D cameras, a growing number of local 3D descriptors that could be used for robotic manipulation tasks have been proposed. However, they are mostly targeted to rigid objects and unstructured point clouds, which makes them computationally expensive, especially when the descriptors need to be densely computed.

Classical 3D descriptors were mostly applied to synthetic CAD models [10], [11] and worked on unstructured point clouds. One of the earliest related approaches, also evaluated on synthetic range images, was proposed in [12], where a histogram representation of a complete point cloud was built merging information of the pixel depths, normal orientations and surface curvatures. More recently, Flint et al. [13] defined a descriptor for Hessian-based interest points by accumulating in a histogram the elevation difference of the normals estimated with different sized planes for all points in a support region.

Given an unstructured point cloud, the Normal Aligned Radial Feature (NARF) descriptor [6] first computes a normal aligned range image centered on an interest point, where points on a local neighborhood are projected onto a plane along the normal direction. The descriptor is then built according to the variation occurring along a number of rays overlapped onto these images. View-point independence is achieved by normalizing to a canonical orientation.

Following a similar philosophy as the SIFT descriptor for 2D images, Tombari et al. [7] presented the Signature of Histograms of OrienTations (SHOT). Given an interest point, the 3D space at its surrounding is split into a fixed number of regions, and the descriptor is built based on histograms of differences between the normal of the points within the region and the normal of the interest point.

Rusu et al. introduced the Point Feature Histogram (PFH) [8] descriptor. It is based on four angle relations computed between every pair of points in a $k$-neighborhood. Each relation is accumulated in a 16-dimensional histogram, yielding a descriptor which is shown to be invariant to position, orientation and point cloud density. Yet, the cost of computing $n$ descriptors on a point cloud is $O(n \cdot k^2)$. Subsequently, the same authors proposed the Fast Point Feature Histograms (FPFH) [9], where instead of computing the relation between every pair of points in a neighborhood, they only considered the connection between the point of interest and its neighbors, and re-weighted the result with descriptor information from the surrounding points. This reduced the cost to $O(n \cdot k)$. Despite being faster, the cost to compute a single FPFH is still large for real-time applications, and it is not applicable to very dense point clouds or situations where one might want to compute descriptors covering a large area.

Summarizing, most of the previous descriptors are designed to work in unstructured point clouds, which translates in that a significant part of the computational cost is associated to defining the neighborhood in which the descriptor has to be computed. This makes that most of them are just computed for a few points of interest. Note, however, that computing reliable 3D interest points might be difficult to accomplish, especially when dealing with 3D data, such as clothes, that do not contain sharp edges.

In addition, in order to gain invariance to viewpoint, existing approaches usually define a local reference frame for each descriptor, which makes re-using computations for descriptors in nearby locations more complicated and introduces an additional cost. However, in our scenario we do not face significant viewpoint changes, but rather simple one-dimensional orientation changes.

Grasping of a textile with a dexterous hand can be performed by aligning the palm to the main direction of the principal wrinkle. Different hand orientations are required for performing the grasp, depending on the orientation of the wrinkle. This can be resolved by having a unique grasp parametrized with the local frame, but also as different hand orientations linked to the different perceptions of the wrinkle, as it is shown in [14], thus allowing a faster descriptor that recycles computations to be used. Furthermore, in principle nothing prevents extending this approach to cover even larger viewpoint areas.

## III. The FINDDD Descriptor

We propose a local descriptor for range images or structured point clouds that takes advantage of the specificities of a clothes manipulation scenario to be both highly discriminative and very fast to compute. First, since the range of depth values within the object will be very limited, we can safely assume that the density of points will be approximately constant over the whole cloud. Next, given that all points in a structured point cloud are distributed in an equally spaced grid or image-like organization, adjacency is well defined. Finally, given that the perceptions of the robot (and the actions performed as a consequence of these perceptions) are intrinsically tied to its current pose, we do not need a specific reference frame for each descriptor, typically used to ensure invariance to rotations.

The combination of the above assumptions allow for very fast computations using integral images. We believe that for a scenario such as robotic manipulation of textile objects, a rapid perception cycle may be more relevant than highly discriminative but very expensive descriptors. Figure 2 summarizes the steps of the method, that will be explained in more detail in the following paragraphs.

Taking inspiration from the SIFT descriptor [16], we define the FINDDD as the concatenation of normal vector (i.e. surface) orientation histograms for several sub-regions inside a support area around the point of interest. The steps to compute the descriptor can be summarized as follows:

1) The normal vector of every point in the input structured point cloud is computed. This step can be made very fast using integral images.
2) For each normal computed in the previous step, we construct a vector of size $B$ with the votes based on the distance between the normal and the neighboring orientation bin centers.
3) To compute a descriptor, all the previously computed vectors within a sub-region are added into an orientation histogram. Then, the orientation histograms of all sub-regions in the support area are concatenated to form the final descriptor vector.
4) Finally, each descriptor is normalized using the L1 norm to make it robust to different densities in the number of points (i.e. NaN points, caused by occlusions or noise, are discarded). Like in the case of SHOT [7], we found it beneficial to keep local density information by only normalizing at the global level.

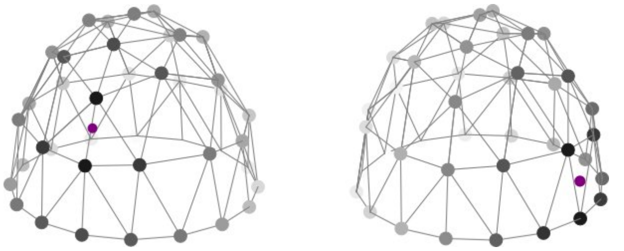


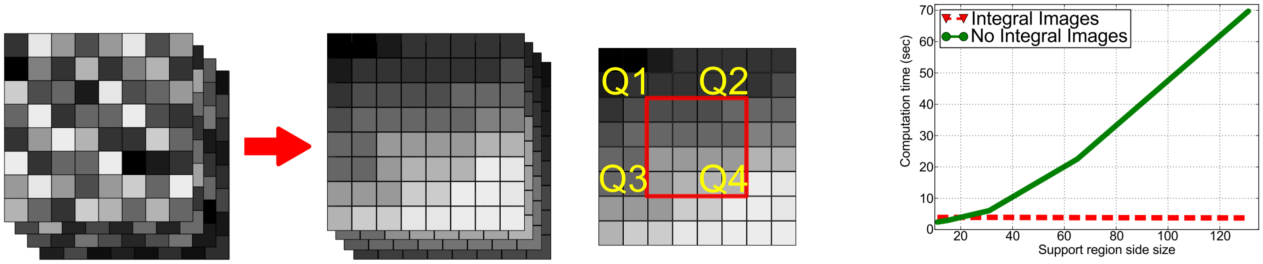
Fig. 3: Detail of the orientation bins with two example vote patterns. The purple dot indicates the coordinates of the normals. 2D voting is usually performed using a grid. The proposed 3D descriptor uses a parametrized division of the sphere to create the voting bins to avoid the typical meridian division that over represents the zone near the north pole.

### A. Orientation Assignment

In order to re-use previously computed data, and in contrast with other works that also use point normal information [7], we do not accumulate the angle between the normal of every point and the central point of the descriptor area. Furthermore, since we are dealing with 2.5D data, only half of the sphere of possible normal orientations has to be considered, which further reduces the size and computational cost of the descriptor.

Since the normal vectors are circumscribed to the unit sphere, a common strategy is to express them as angles in spherical coordinates (e.g. Hetzel et al [12]). However, defining the orientation bins in the angular space has some caveats: first, bins do not cover the same sphere space in all locations, and are more concentrated around the north pole (maximum elevation), which leads to an irregular representation of the normals; second, azimuth information becomes unstable as we get closer to the maximum elevation point, and small changes due to noise can easily produce swaps in the assigned bin.

Instead, we define bins distributed across the entire semi-sphere in Cartesian coordinates. Precisely, we use the vertex points generated in a triangular tessellation to obtain a quasi-regular distribution of the orientation centers (see Figure 3). One alternative yielding completely regular bins is the ap-



(a)

(b)

Fig. 4: (a) Integral images permit computing the sum of any image sub-rectangle with only four basic operations. Keeping one integral image for each bin of our histogram makes descriptor extraction linear in the number of descriptors. (b) Comparison of computation time for descriptors extracted densely (every pixel), varying the size of the local region used.

proach of Klaser et al. [15], where points in the sphere surface are projected onto a platonic solid; however, it has a limitation on the number of bins, since the platonic solid with more facets available is the icosahedron (20-sided).

The downside of our representation is a higher cost to assign a normal to its corresponding bin, because the distance to the bin centers in the unit half-sphere surface has to be computed. However, for a reasonably small number of bins, there is no noticeable slow-down in computation and, if a larger number of bins is desired (e.g. for data coming from a very precise 3D sensor), structures like K-D Trees can significantly accelerate the search.

Increasing the number of orientation bins of the descriptor improves the angle resolution of the model, and consequently the accuracy with which surfaces can be represented. However it also induces aliasing, and sparsity, which degrade the significance of distances between descriptors. Another consideration regarding the number of bins is that it must be adjusted to the level of noise inherent in the input data, which may otherwise worsen the aliasing problem.

We mitigate these two problems using soft voting to interpolate between different bins. Moreover, the distribution of the orientations of the normals can be more accurately captured in this way, and that has been shown to increase the robustness of local descriptors by avoiding the aliasing problems of hard assignment [16]. Precisely, we allow each normal to contribute to all bins closer than a unit of histogram bin spacing.

### B. *Efficient Computation Using Integral Images*

As mentioned earlier, by using structured point clouds, it is possible to take advantage of integral images[1], and make the computational cost linear in the final number of descriptors: only $O(4 \cdot o \cdot n \cdot s)$ operations, where $s$ is the fixed number of spatial sub-divisions (typically 16, for a $4 \times 4$ grid), $o$ is the number of orientation bins we are considering and $n$ the number of descriptors to extract from the point cloud. We use integral images both to compute the normal vectors[2] and to perform the aggregation of the votes for every orientation bin and sub-region of the descriptor.

An integral image [17] is a representation of the integral of the pixel values, and is computed in the following way:

$$I(i,j) = M(i,j) + I(i-1,j) + I(i,j-1) - I(i-1,j-1),$$

where $I$ is the integral image, $M$ is the original image we want to perform summations on, and $i$ and $j$ are sub-indexes that iterate all the rows and columns of the image. If one sub-index becomes negative, the term is removed. Then, the orientation histogram for any sub-region of the image can be computed by performing only four basic operations for each orientation bin (see Figure 4a):

$$\Sigma = Q1 + Q2 - Q3 - Q4,$$

where $\Sigma$ is the sum of all the pixels inside the rectangle in Figure 4a, and Q1 to Q4 are the values of the integral image at their given position in the figure. In our case, to compute the descriptors, one integral image is necessary for each orientation bin, where we will accumulate the occurrences of normals falling into it from the top-left to the bottom-right corner of the structured point cloud. Consequently, the cost of constructing the integral images is $O(o \cdot p)$, where $p$ is the total number of points in the point cloud.

Figure 4b shows a comparison of the computational time required to extract a descriptor for every point of a $640 \times 480$ structured point cloud using a 3Ghz Linux machine. Note that in the case where integral images are not used, the neighborhood information provided by the structured point cloud is still used and, in an unstructured point cloud, we would require an extra step of searching for the nearest neighbors of each point in which we want to compute a descriptor. This would require typically using a K-D tree for nearest neighbor search, at an additional cost of $O(log(q))$ for each descriptor, where $q$ is the number of points in the point cloud, plus $O(q \cdot log(q))$ for building the tree once.

[1] This is the name by which the technique of summed-area tables has become popular in the computer vision community, and is used here for this reason. This technique is in fact applicable to any matrix-like object (structured point clouds in our case).

[2] To compute the normal vectors using integral images, we use the implementation in the PCL library.

TABLE I: Comparison of the computational cost of the evaluated descriptors on a 3Ghz Linux machine. Result obtained extracting a descriptor for every point in a $640 \times 480$ structured point cloud (average of 10 runs). Parameter B for FINDDD stands for the number of orientation bins considered. Column **Time** shows the time spent computing the descriptors for the whole point cloud, and column **Time per desc** the average time for a single descriptor.

| Descriptor | Time (s) | Time per desc. (ms) |
|---|---|---|
| FINDDD (B=13) | 4.0 | 0.0153 |
| FINDDD (B=41) | 10.5 | 0.0402 |
| SHOT | 482.5 | 1.98 |
| FPFH | 313.5 | 1.28 |

## IV. Experimental Results

To evaluate the performance of the FINDDD, we performed four tests: An efficiency assessment where we compared against state-of-the-art descriptors, a wrinkle retrieval experiment, a classification of garments made of different textile materials, and an informed grasping point selection.

### A. *Efficiency Assessment*

To evaluate the computational cost of our descriptor, we measured the time necessary to extract a descriptor for every point of a $640 \times 480$ point cloud acquired with a Kinect range camera. Points at the edge of the point cloud, without enough neighbors for the support region, were discarded. Using more orientation bins increases the number of integral images that have to be constructed and the operations to compute each sub-region. Consequently, we have tested the method with 13 and 41 orientation bins. These will be the same values we will use in the recognition experiments we report in the following sections.

Two other state-of-the-art descriptors, SHOT and FPFH[3], are also evaluated to provide a reference to our results. The setup is the same as the one used for our descriptor, but in this case points with a NaN value in any coordinate were filtered out from the input point cloud as advised in the documentation of the descriptors.

Table I shows the results of the comparison. Note that FINDDD, in both its configurations, is two orders of magnitude faster than both FPFH and SHOT. We will next show that the recognition rates when characterizing textiles, are very similar for all methods.

### B. *Wrinkle Retrieval*

We perform an initial evaluation of the proposed descriptor performance on an in-house dataset containing $640 \times 480$ Kinect RGB-D images of a polo shirt showing one of eight distinct manually produced wrinkles (see Figure 5). Five repetitions of each wrinkle were acquired, and the relevant wrinkle area in each image was annotated by hand. Then, we extracted pixel-wise descriptors for each image, and selected the center of gravity of the annotation as the representative for the particular wrinkle and image. We also stored a fixed number of descriptors from random points in the annotated regions for additional testing.

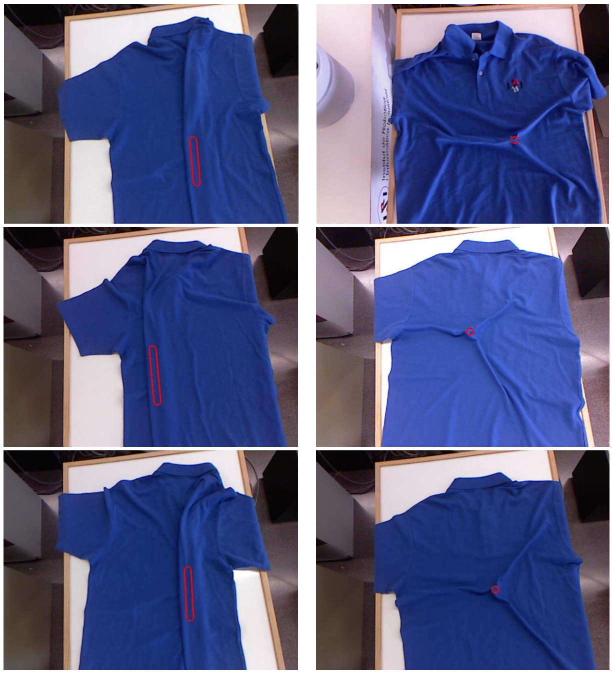


Fig. 5: Samples of two wrinkles from the dataset; three repetitions of each one are shown (column-wise). The annotated wrinkle area can be seen encircled in red.

Using this dataset we evaluate the retrieval performance of FINDDD. The distance to a query descriptor is used to re-order all the descriptors in the dataset, and the average precision of the resulting list is computed. The test is repeated for every instance of every wrinkle type, and the mean average precision is reported. The same queries are performed in two databases: one where only the representative descriptors are present, and another that additionally contains the descriptors from all the previously selected random points, labeled according to the instance they are drawn from. We also evaluated the performance of FPFH and the SHOT descriptors in the same way. The results of this test can be seen in Table II.

From the results in the table it can be seen that, in our textile manipulation setup, our proposed descriptor is able to correctly characterize and recognize wrinkles with a performance similar or even superior to that of state-of-the-art descriptors like FPFH or SHOT. It is also noticeable that being able to use a large enough support region is essential to properly characterize a textile wrinkle, which our descriptor is able to do at no additional computational cost. Finally, we see that when 41 orientation bins are used, aliasing causes some degradation in the results, but also that soft voting compensates for this and is able to maintain the performance.

### C. *Garment Recognition*

We also compared the proposed descriptor with FPFH and SHOT in a class recognition task consisting in distinguishing

[3] Both descriptors were computed using the Point Cloud Library (PCL) 1.5, and with the parameters suggested in the tutorials of the library.

TABLE II: Results of a wrinkle retrieval experiment. **B** stands for the number of orientation bins used, **S** for the side (in pixels) of the support region, and **SV** shows if soft voting was used or not (True/False). Column **Rep** shows the mean average precision of the test where only the representative points were used, while column **Ext** shows the results for the test that included the additional random points.

| Descriptor | B | S | SV | Rep | Ext |
|---|---|---|---|---|---|
| FINDDD | 13 | 21 | T | 48.9 | 59.8 |
| | 13 | 21 | F | 45.6 | 52.2 |
| | 13 | 43 | T | 66.2 | 80.5 |
| | 13 | 43 | F | 64.6 | 80.6 |
| | 13 | 65 | T | 68.2 | 85.9 |
| | 13 | 65 | F | **68.9** | **86.1** |
| | 41 | 21 | T | 48.8 | 61.2 |
| | 41 | 21 | F | 46.6 | 57.1 |
| | 41 | 43 | T | 66.8 | 82.2 |
| | 41 | 43 | F | 61.8 | 77.6 |
| | 41 | 65 | T | 68.8 | 86.0 |
| | 41 | 65 | F | 64.3 | 81.8 |
| FPFH | | | | 47.2 | 62.0 |
| SHOT | | | | 40.2 | 50.6 |

clothes made of different types of textile materials based on the wrinkles they produce. We used the publicly available *IRI Clothing Part* dataset[4], that contains RGB-D images of six classes of garments.

We split the dataset in two parts (70% train and 30% test), and represented the images using bag of features (BOF) models of size 512. Then we trained a Support Vector Machine (SVM) classifier with either linear or RBF-$\chi^2$ kernel on the BOF for each textile material class. The average precision obtained by FINDDD (13 orientations and 43-pixel sided support regions), FPFH and SHOT descriptors in this test can be seen in Table III. As can be observed, the performance of FINDDD is comparable to that of the two state-of-the-art descriptors. The superior performance with the polo is attributable to the higher number of training samples for that category.

Note that this test is not directly comparable to the recent work by Willimon et al. [18], since they are doing clothing classification with a dataset focused on intra-class variation using a sophisticated approach containing both appearance and depth information, as well as mid-level feature information, while here we are distinguishing between different exemplar garments using only depth information.

### *D. Informed Grasping Point Selection*

Finally, we used the FINDDD descriptor to refine the final grasping location in the polo collar detection pipeline of Ramisa et al. [4]. The objective is to ensure that the polo shirt is grasped from the lapel of the collar, so it can be automatically hanged on a clothes hook hanger.

The collar detection method predicts bounding boxes that are likely to contain the collar of a polo shirt, but a generic "graspability" measure is used to determine the final grasping point within the bounding box [4]. On the contrary, the task of hanging the polo on a hook hanger is greatly simplified if the robot is able to grasp the polo by the lapel. In this experiment we use our proposed descriptor to automatically select a grasping point on the lapel of the polo using a RBF-$\chi^2$ SVM classifier trained on descriptors from a few polo images conveniently annotated as lapel/non-lapel. The test was attempted 30 times, and from the cases where the collar was correctly detected (76%), a point in the lapel could be correctly located 82.6% of the times. Some positive and negative examples can be seen in Figure 6.

TABLE III: Average precision obtained by the different descriptors in the garment recognition task.

| | Linear SVM | | | RBF-$\chi^2$ SVM | | |
|---|---|---|---|---|---|---|
| Garment | FINDDD | SHOT | FPFH | FINDDD | SHOT | FPFH |
| Dress | 37.5 | 25.5 | **51.2** | 66.8 | 61.9 | **67.6** |
| Shirt | 41.7 | 41.2 | **45.1** | 54.5 | 72.9 | **79.7** |
| T-Shirt | **71.8** | 58.1 | 68.9 | **84.7** | 70.1 | 76.5 |
| Jeans | 41.5 | 33.4 | **58.1** | 72.9 | 65.1 | **77.9** |
| Polo | **85.2** | 64.7 | 71.6 | **96.0** | 83.7 | 77.6 |
| Sweater | 44.8 | 36.5 | **80.1** | 84.6 | 92.1 | **93.7** |
| Average | 53.7 | 43.2 | **62.5** | 76.6 | 74.3 | **78.8** |

## V. Conclusions and Future Work

In this work we have presented a novel shape descriptor for range images and structured point clouds, designed for robotic manipulation of textile objects, that trades invariance to point of view for very fast computation thanks to the use of integral images.

The proposed descriptor has been evaluated through different tests, yielding good results and with a performance comparable or even better than other state-of-the-art 3D descriptors in the particular domain of perception for robotic manipulation of textile objects.

Regarding future work we plan to evaluate the FINDDD descriptor within a robotic manipulation pipeline. In addition, we are investigating multiple-cue tracking methodologies [19], [20] to merge the proposed geometry-based descriptor with appearance-based ones, especially with those which are specifically designed to handle non-rigid deformations [21] and occlusions [22], situations typically encountered when manipulating clothes.

## References

[1] 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 *Proc. International Conference on Robotics and Automation* , 2010, pp. 2308–2315.

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

[3] S. Miller, M. Fritz, T. Darrell, and P. Abbeel, "Parametrized Shape Models for Clothing," in *Proc. International Conference on Robotics and Automation* , 2011, pp. 4861–4868.

[4] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *Proc. International Conference on Robotics and Automation* , 2012, pp. 1703–1708.

[5] B. Willimon, S. Birchfield, and I. Walker, "Classification of Clothing using Interactive Perception," in *Proc. International Conference on Robotics and Automation* , 2011, pp. 1862–1868.

[4] `http://www.iri.upc.edu/groups/perception/clothing_dataset/`

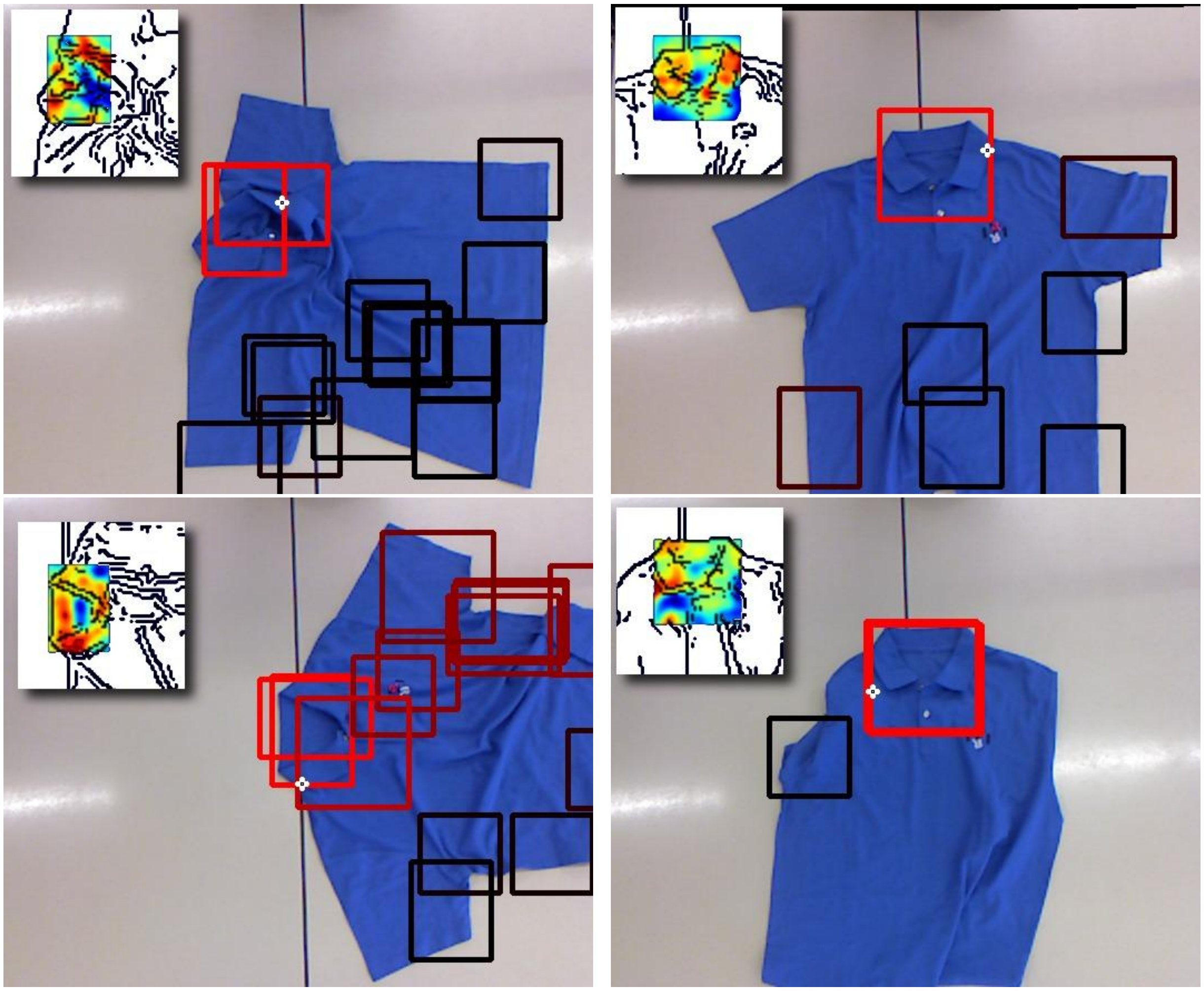


Fig. 6: Sample results for the lapel classification experiments (best viewed in color). The bounding boxes correspond to the collars detected by [4] (normalized from red to black). Overlaid on the top left of every image, there is the score map obtained by the SVM lapel classifier within the highest ranked bounding box (red means higher and blue lower), with Canny edges superimposed for clarity. The final grasping point (highest score from SVM) is marked with a white cross.

[6] B. Steder, R. Rusu, K. Konolige, and W. Burgard, "Point feature extraction on 3d range scans taking into account object boundaries," in *Proc. International Conference on Robotics and Automation* , 2011, pp. 2601–2608.

[7] F. Tombari, S. Salti, and L. D. Stefano, "Unique signatures of histograms for local surface description," in *Proc. European Conference on Computer Vision*, 2010, pp. 356–369.

[8] R. Rusu, Z. Marton, N. Blodow, and M. Beetz, "Persistent point feature histograms for 3D point clouds," in *Intelligent Autonomous Systems 10* , 2008, p. 119.

[9] R. B. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3D registration," in *Proc. International Conference on Robotics and Automation* , 2009, pp. 3212–3217.

[10] F. Stein and G. Medioni, "Structural indexing: efcient 3-D object recognition," *IEEE Trans. Pat. An. Mach. Int.*, vol. 14, no. 2, pp. 125–145, 1992.

[11] A. E. Johnson, "Spin-Images: A Representation for 3-D Surface Matching," Ph.D. dissertation, Carnegie Mellon University, 1997.

[12] G.Hetzel, B.Leibe, P.Levi, and B.Schiele,"3D object recognition from range images using local feature histograms,"*Proc. Conf. on Comp. Vision and Pattern Recognition*, pp.2:394–399, 2001.

[13] A. Flint, A. Dick, and A. V. D. Hengel, "Thrift : Local 3D Structure Recognition," in *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society*, 2007, pp. 182–188.

[14] G. Alenyà, A. Ramisa, F. Moreno-Noguer, and C. Torras, "Characterization of Textile Grasping Experiments," in *Workshop on the Conditions for Replicable Experiments and Performance Comparison in Robotics Research (in conjunction with ICRA 2012)*, 2012.

[15] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *19th British Machine Vision Conference* , 2008, pp. 275:1–10.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] F. C. Crow, "Summed-area tables for texture mapping," in *ACM SIGGRAPH Computer Graphics*, vol. 18, no. 3, 1984, pp. 207–212.

[18] B. Willimon, I. Walker, and S. Birchfield, "A new approach to clothing classification using mid-level layers," in *Proc. International Conference on Robotics and Automation*, 2013, pp. 4256–4263.

[19] F. Moreno-Noguer, J. Andrade-cetto, and A. Sanfeliu, "Fusion of color and shape for object tracking under varying illumination," in *Proc.IBPRIA, LNCS 2652*. Springer, 2003, pp. 580–588.

[20] F. Moreno-Noguer, A. Sanfeliu, and D. Samaras, "Integration of deformable contours and a multiple hypotheses fisher color model for robust tracking in varying illuminant environments," *Image Vision Comput.*, vol. 25, no. 3, pp. 285–296, Mar. 2007.

[21] F. Moreno-Noguer, "Deformation and illumination invariant feature point descriptor," in *Proc. Conference on Computer Vision and Pattern Recognition* , 2011, pp. 1593–1600.

[22] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer, "Dense segmentation-aware descriptors," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2013.