Engineering Applications of Artificial Intelligence ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Learning RGB-D descriptors of garment parts for informed robot grasping

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

Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Llorens i Artigas 4-6, 08028 Barcelona, Spain

ARTICLE INFO

Article history:
Received 16 August 2013
Received in revised form
26 May 2014
Accepted 27 June 2014

Keywords:
Computer vision
Pattern recognition
Machine learning
Garment part detection
Classification
Bag of Visual Words

ABSTRACT

Robotic handling of textile objects in household environments is an emerging application that has recently received considerable attention thanks to the development of domestic robots. Most current approaches follow a multiple re-grasp strategy for this purpose, in which clothes are sequentially grasped from different points until one of them yields a desired configuration.

In this work we propose a vision-based method, built on the Bag of Visual Words approach, that combines appearance and 3D information to detect parts suitable for grasping in clothes, even when they are highly wrinkled.

We also contribute a new, annotated, garment part dataset that can be used for benchmarking classification, part detection, and segmentation algorithms. The dataset is used to evaluate our approach and several state-of-the-art 3D descriptors for the task of garment part detection. Results indicate that appearance is a reliable source of information, but that augmenting it with 3D information can help the method perform better with new clothing items.

1. Introduction

Handling highly flexible objects, such as clothes, is a robotic application that is attracting increasing attention. It is a challenging task since the high-dimensional configuration space of a textile object makes it very difficult to determine its state and, consequently, plan the appropriate actions to bring it into a desired configuration. Also, the success of manipulation actions becomes very unpredictable.

In this work, we propose to use machine learning to detect clothing parts as a first step towards informed grasping of garments. Our approach is based on the well-known Bag of Visual Words (BoVW) (Csurka et al., 2004) method from computer vision. BoVW seems better suited to cope with the challenges of detecting parts in flexible objects since it does not impose a geometry as template matching or pictorial structure methods do (Parkhi et al., 2011). Our objective is to assess if such technique can be used to recognize clothing parts and with what accuracy. Moreover we seek to demonstrate the intuition that adding 3D information should improve the results compared to only appearance information.

The first contribution is an evaluation of our proposed approach for different garment part detection and classification tasks, combining SIFT with four state-of-the-art 3D descriptors. In order to evaluate the method we have collected and manually annotated a large dataset of RGB-D scans of clothing items. The publication of such dataset constitutes the second contribution of this work, and promotes further comparison and benchmarking. The dataset is intended for research on detecting garment parts under severe deformations, not on classifying garment types with very different appearance, i.e. there are tens of images of only two t-shirts under a broad range of deformations instead of lots of different t-shirts.

This work is an expanded version of Ramisa et al. (2012). The original approach is modified to be faster at test time, and is evaluated much more thoroughly: the main dataset used for evaluation is greatly extended, and additional garment parts are tested with several shape descriptors. Furthermore, a second, independent, dataset is included in the evaluation.

2. Related work

In this section we will briefly review related work in the three areas addressed by this paper: 3D shape descriptors, current approaches to garment perception and manipulation and, finally, existing datasets for the evaluation of methods related to the perception of garments.

3D descriptors: The recent availability of consumer level 3D imaging devices based on structured light (SL), such as the Kinect,

* Corresponding author.

http://dx.doi.org/10.1016/j.engappai.2014.06.025
0952-1976/

has boosted research on 3D perception, which is experiencing a surge of new descriptors and techniques, as image-based perception experienced a decade ago. 3D perception plays a significant role in robotics, being one of the most pressing bottlenecks for the wide spread of robotic solutions to everyday problems. Economic and reliable 3D sensors offer a wealth of new opportunities to develop useful robotic applications for domestic environments.

Garments may have very different color combinations, designs and imprints, but may also have only one color. Furthermore, their appearance can change a lot due to severe deformations. Consequently, classical appearance descriptors might have a poor performance, and thus the 3D information provided by RGB-D cameras should apparently provide a key advantage. Additionally, the ability to supply registered color and depth allows us to design new descriptors based on the depth and color together.

Appearance-based descriptors, such as SIFT (Lowe, 2004), have been used for more than a decade, and throughout this time, they have been carefully engineered to produce the highest quality possible results. Unfortunately, this is not yet true for 3D descriptors, which have, until recently, attracted little attention from the computer vision and robotics communities, mainly because of the scarcity, drawbacks and cost of previous 3D imaging devices.

Early work on 3D descriptors focused on areas such as CAD model retrieval, where no perception from the environment was involved, and models were always complete (Tangelder and Veltkamp, 2004). Conversely, works in other areas such as simultaneous localization and mapping (SLAM) used expensive 3D sensors like LIDARs to acquire point clouds from the environment, but most of the focus was on multiple scan registration to construct large maps, not until recently feature extraction from LIDAR data has attracted significant attention of the research community (Himmelsbach et al., 2009; Li and Olson, 2010). In general, previous efforts in 3D descriptor research have mainly concentrated on the case of rigid objects (Lai et al., 2011; Janoch et al., 2011) or, at most, articulated objects (Shotton et al., 2011). To our knowledge, this is the first work evaluating different 3D descriptors for garment part recognition.

Garment perception and manipulation: Although there exist a wide literature on the perception of deformable cloths using only RGB information (Moreno-Noguer et al., 2009; Sanchez et al., 2010; Moreno-Noguer and Fua, 2011), most of these approaches are computationally expensive, and not ready to be used in real garment manipulation settings. For this purpose, robotic applications resort to the use of RGB-D sensors. Several recent works have addressed this task with limited, although encouraging, results. Maitin-Shepard et al. (2010) present a system that can fold one by one all elements on a pile of washed towels. Vision is used to detect the corners of the towel that a PR2 robot is holding, and it is re-grasped until a desired state has been reached. From this known state the towel is folded in an open-loop procedure. Later, Cusumano-Towner et al. (2011) describe an improved, end-to-end, laundry handling system. A garment is picked up, identified, and brought into a desired configuration. In order to carry out the task, a series of re-grasps are executed by the robot, and a Hidden Markov Model and a clothing simulator are used to identify the item and its pose, as well as to plan how to bring it into the desired state. An interactive vision system is proposed by Willimon et al. (2011) that iteratively identifies all items in a basket of clothes using four basic visual features and the 1-nearest neighbor classifier. Wang et al. (2011) propose a system for the manipulation of socks with a PR2 robot that uses state-of-the-art computer vision techniques. Willimon et al. (2013) present a system to determine the type of clothing items laying on a flat surface using a variety of low-level appearance and depth features, as well as mid-level layers such as attributes (e.g. stripped, v-neck, front zipper) of the garments. These works show a trend towards the usage of more sophisticated perception techniques in robotic clothing handling applications, as well as a pervasive use of 3D information.

Garment datasets: Because of its complexity, perception of garments is a field that has just recently been undertaken by the research community, and consequently there is a lack of well-established benchmarks for its multiple tasks and applications.

One such task that has received some attention, partly because of its significance in surveillance applications, is identifying the garments worn by people in pictures or videos, but there the focus is on detecting the pose of the people and on the appearance of clothes when being worn, as opposed to being centered on the garments themselves, and on the large space of states they can adopt. Datasets in this category include Yamaguchi et al. (2012) Fashionista dataset and, closer to our problem as it focuses on the clothes themselves, the dataset proposed by Hidayati et al. (2012). Unfortunately these datasets do not include depth information or garment part annotations.

There are also some datasets more focused on robotic applications, as the ones proposed by Yamazaki and Inaba (2013) and by Mariolis and Malassiotis (2013), that deal with classifying (possibly wrinkled or folded) garments laying on a flat surface, using only appearance information. None of them includes depth data or garment part annotations either.

Doumanoglou et al. (2014) propose a dataset of six clothes hanging from a gripper. It includes depth information, but no annotation of parts. Besides, the way in which the garments are presented to the camera, makes it very difficult that the parts we are interested in are visible in the image.

Other datasets (Aragon-Camarasa et al., 2013; Willimon et al., 2013) present the clothes laying on a flat surface and do include depth data, but are focused on tasks like stereo depth estimation, classification or folding/flattening of clothes, so no part annotations are included.

Finally, more related to ours is the very recent CTU Color and Depth Image Dataset of Spread Garments (Wagner et al., 2013) that includes appearance and depth data, as well as annotations that, despite not being designed for our tasks, are amenable to it. We conducted some additional experiments on this dataset to further evaluate the proposed method.

3. Garment part detection method

As said in the Introduction, the long-term goal of this research is to perform informed graspings, which can be useful for an end-to-end clothing handling system like the one of Cusumano-Towner et al. (2011), for example to shorten the series of re-grasps necessary to verify that the clothing is in a desired configuration. We attempt to use state-of-the-art computer vision techniques to detect the relevant grasping points from the very beginning, while the object is still laying on the table/surface. For this we propose a vision and depth based detection method, consisting of a coarse-to-fine architecture based on the well-known "Bag of Visual Words" (BoVW) (Csurka et al., 2004) image representation, widely used in the computer vision literature, and a sliding window approach. A schema of the proposed method can be seen in Fig. 2. Here, we are not performing robotic grasping experiments, hence we are not using the *grasping point selection* step proposed by Ramisa et al. (2012). At this stage, and as done in related work, we are not considering the problem of background subtraction as a significant body of work is already addressing it (e.g. Felzenszwalb and Huttenlocher, 2004; Yang et al., 2012; Rashedi and Nezamabadi-pour, 2013; Grady, 2006). We assume a segmentation method able to precisely select the garment is available.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

3.1. Appearance and depth local features

Our detection method is based on the appearance and depth information, obtained from the Kinect image. Both types of features are quantized using visual vocabularies learned with K-means from a large training database of descriptors. A BoVW descriptor can be then constructed by accumulating in a histogram all the visual words present in a local neighborhood defined by a bounding box (see Fig. 1). Combinations of two descriptors are then formed by concatenating the two BoVW vectors.

In order to obtain the appearance information, we use the well-known scale invariant feature transform (SIFT). This descriptor divides a local patch around the interest point in 16 sub-regions, and computes a 8-bin histogram of the orientations of the gradient for each sub-region, weighted by its corresponding magnitude and a Gaussian applied at the center of the patch. In order to reduce the aliasing in the orientation, trilinear interpolation is used to distribute gradient samples across adjacent bins of the histograms. Next, the histograms are concatenated, yielding a 128-dimensional descriptor. To reduce the influence of non-affine illumination changes, the normalized descriptor is thresholded at 0.2 and re-normalized.

Regarding the depth information, we evaluate several recently proposed 3D descriptors: the Geodesic-Depth Histogram (GDH), the Fast Point Feature Histogram (FPFH) (Rusu et al., 2009), the Heat Kernel Signature (HKS) (Sun et al., 2009) and the Fast Integral Normal 3D (FINDDD) descriptor (Ramisa et al., 2013). A short description of the four depth descriptors follows.

GDH: The Geodesic-Depth Histogram captures the joint distribution of geodesic distances and depths within the patch. It is an adaptation to depth information of the Geodesic-Intensity Histogram, originally introduced by Ling and Jacobs (2005) for describing deformable image patches.

Let us consider a patch $\mathcal{P}$ in the image, centered on a point of interest p, that in our case corresponds to every point of a grid that densely covers the image. Each point $p_i \in \mathcal{P}$ has an associated depth value d_i obtained from the Kinect camera. Then the histogram for p is computed as follows:

- The histogram is initialized by splitting the joint space of geodesic distance and depth into a discrete number of intervals.
- For each $p_i \in \mathcal{P}$, the geodesic distance g_i is computed with respect to p, using the Fast Marching algorithm (Sethian, 1996).
- Then the bins of the histogram are filled with each pair (d_i, g_i) of depth and geodesic distance values.

The descriptor of p is finally built by concatenating the value of all the bins in the histogram.

FPFH: The Fast Point Feature Histogram descriptor (Rusu et al., 2009) (a simplification of the Point Feature Histogram descriptor (Rusu et al., 2008)) is designed to characterize the local geometry around a point in a point cloud. Given a point p_q for which we want to compute the descriptor, for each of its k-nearest neighbors, a local coordinate frame $\langle u, v, w \rangle$ between the query point and its neighbor p_t is determined, and three geometrical relations are computed:

$$\alpha = v \cdot n_t \tag{1}$$

$$\phi = u \cdot \frac{(p_t - p_q)}{d} \tag{2}$$

$$\theta = \arctan(w \cdot n_t, u \cdot n_t) \tag{3}$$

where d is the Euclidean distance between the points p_q and p_t, and n_q and n_t are the normals at the two points in the local coordinate frame.

Then, a similar descriptor is computed for each of the k neighbors, in their own k-neighborhood, and a weighted sum of the simplified descriptors is performed to incorporate all the information in the final FPFH descriptor.

HKS: The Heat Kernel Signature (Sun et al., 2009) is a shape descriptor based on the heat diffusion equation applied to a shape modeled as a Riemannian manifold that has been shown to give good results in non-rigid 3D shape recognition. Later, the descriptor has been made scale-invariant by using its Fourier transform and a logarithmic sampling (Bronstein and Kokkinos, 2010). It has also been shown to work well using photometric information (Moreno-Noguer, 2011).

Put in simple terms, this descriptor models the evolution of the temperature of the nodes in a mesh after an input of a unit of heat has been applied at a given point. It is motivated by the fact that isometric deformations of the shape that do not change its topology will not change the way the heat is diffused.

To reduce the computational cost, and since we want the descriptor to be local, we first segment a local region centered at the point of interest, and compute the HKS in the segmented mesh. Finally, following Bronstein and Kokkinos (2010), a logarithmic sampling in the time scale and fast Fourier transform of the heat signature are applied in order to obtain a scale invariant representation.

FINDDD: The Fast Integral Normal 3D descriptor (Ramisa et al., 2013) represents the distribution of orientations of the 3D normals in a region around a point of interest in a structured point cloud (e.g. a Kinect scan). Thanks to using integral images, the FINDDD descriptor can be computed densely over a point cloud up to two orders of magnitude faster than FPFH. Spatial subdivisions are incorporated to better represent the area around the point.

The computation of the FINDDD descriptor is done as follows: the 3D normal is computed for every point in the cloud. Then, at each position of a grid defined over the point cloud (as it is

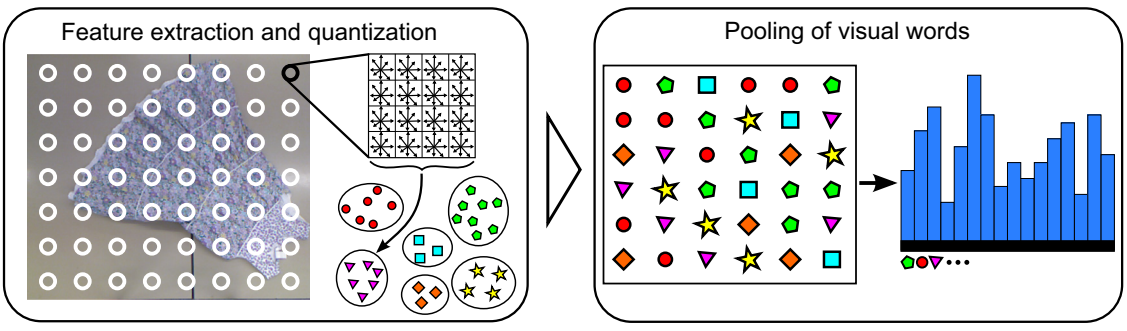

Fig. 1. Steps of the BoVW vector construction. First descriptors are extracted from the appearance or depth image and, next, quantized into visual words using a vocabulary previously learned and pooled in a histogram of visual word counts. This procedure can use the whole image, or be restricted to a region of interest.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

structured, it can be seen as an image or 2D matrix), a descriptor is computed by constructing normal orientation histograms for each sub-region inside the local region.

Instead of using bins defined as angles in spherical coordinates, they are distributed regularly across the entire semi-sphere in Cartesian coordinates. This avoids the singularity at the north pole and the uneven area assigned to each bin caused by the angular representation.

3.2. Detection probability map

With BoVW descriptors constructed from positive and negative training bounding boxes, a logistic regression model is trained using LIBLINEAR (Fan et al., 2008) to obtain the posterior probability of the garment part being present in a given bounding box. The probability of a bounding box containing the part of interest (class C_+) given a BoVW descriptor x can be computed as

$$p(C_+|x) = \frac{1}{1+e^{w^T x}} \tag{4}$$

where w are the parameters of the model, learned minimizing the following expression:

$$\min_w \left(\frac{1}{2} w^T w + C \sum_{i=1}^{N} \log(1+e^{-y_i w^T t_i}) \right) \tag{5}$$

where C is the regularization parameter (adjusted by cross-validation), t_i stands for the ith training example and y_i is its corresponding label.

Positive samples are the annotated bounding boxes in the training set, and negatives are random bounding boxes, sampled from the clothing area, that do not have more than 50% overlap with the annotated bounding box according to the Jaccard index:

$$I_{Jaccard} = \frac{area(B_n \cap B_{gt})}{area(B_n \cup B_{gt})} \tag{6}$$

where B_n is the generated negative bounding box and the B_{gt} is the ground truth one.

In the first layer of the architecture (corresponding to steps b and c of Fig. 2) the logistic regression model is used in a sliding window approach covering the whole image, with different sizes and shapes of bounding boxes drawn from the distribution of those annotated in the training set. In order to accelerate the computation of the sliding window classifier scores, we use the Efficient Histogram-based Sliding Window-Dense approach from Wei and Tao (2010). Next, similarly as it is done in Aldavert et al. (2010), the probabilities of all windows are combined in a probability map of the presence of the garment part. Local peaks of this probability map are then selected and passed to the second layer of the architecture.

3.3. Detection refinement

A linear method like logistic regression has the advantage of being fast to apply at test time, but its performance is sometimes limited. A type of classifier with more capacity, and specifically designed for histograms, is the Support Vector Machine with the χ^2 extended Gaussian kernel (Zhang et al., 2006):

$$\chi^2(x,t) = \exp\left(-\gamma \sum_j \frac{(x_j - t_j)^2}{x_j + t_j} \right) \tag{7}$$

where γ is the inverse of the average of the χ^2 distance between the elements on the training set.

To refine the results of the logistic regressor, in the second layer (corresponding to step d in Fig. 2), for each selected candidate point, we cast a set of windows of different shapes and offsets with respect to the original point. Next, the score assigned by a χ^2 extended Gaussian kernel SVM is used to rank these new windows (we use Platt's probabilistic outputs algorithm (Platt, 1999; Lin et al., 2007) to convert the scores back to probabilities), and only the highest ranked window for each candidate point is accepted. In practice we are conducting a local search around the most probable locations of the part with a more expensive but reliable classifier. The parameters of the χ^2 extended Gaussian kernel SVM are also tuned by cross-validation, but other methods, for example that of Lázaro-Gredilla et al. (2012), could be used.

3.4. Image-Level Prior

In order to increase the precision of the detector, we evaluate the impact of incorporating an Image-Level Prior (ILP) (Shotton et al., 2008), which provides the probability that a given image contains the class of interest. The ILP is also based on a Bag of Visual Words method, but it uses information from the whole image (filtered with the segmentation mask) to learn a logistic regression classifier, which then gives the probability for the presence of the part of interest. If the probability is too low, the detector for that particular class is not applied to the image (see Fig. 3 for an example).

4. Clothing part dataset

In order to test how the different 3D descriptors work for highly flexible objects, we created a dataset of registered images and point clouds, acquired with a Kinect structured-light camera. Each scan shows one or various everyday clothing items laying on a table with parts, such as collar or sleeves, annotated by hand with polygons.

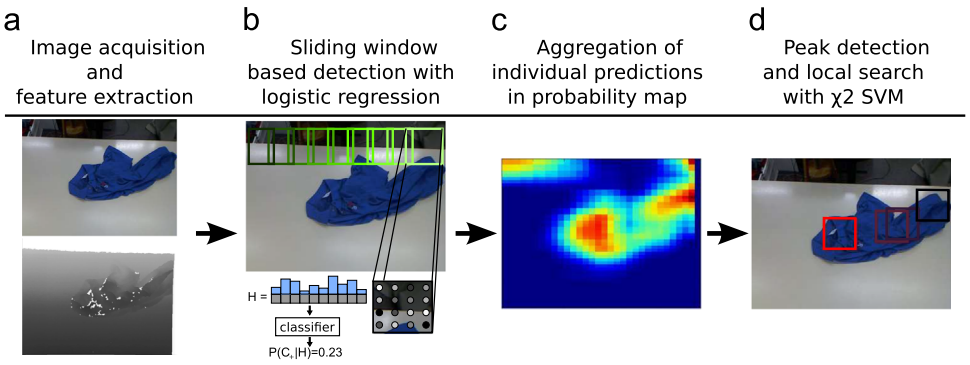

Fig. 2. Schema of the method proposed. Steps b and c correspond to the first layer of the approach as described in the text. Step d corresponds to the second layer, and step e to the third. In the image of step d, reddish color of the bounding box indicates more confidence in the detection. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

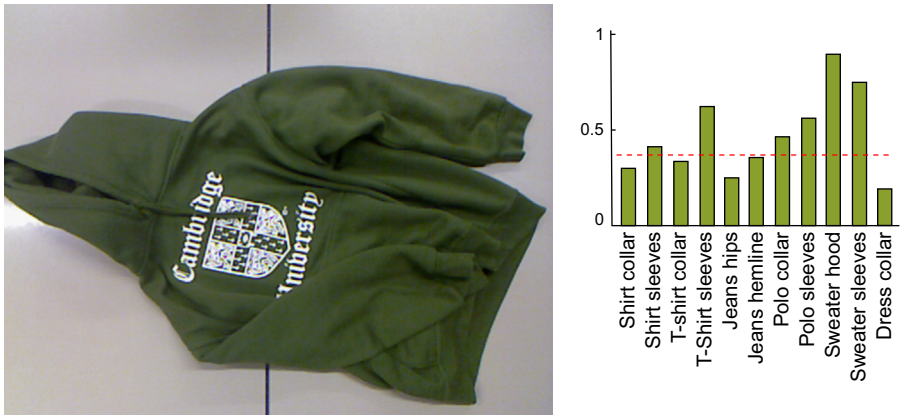

Fig. 3. The Image-Level Prior consists of a probability for each garment part (classes) for the complete image. Classes with a probability below a threshold, represented by a dashed line, are not searched for, saving computation time and avoiding false positives.

Table 1
Criteria used during the annotation process.

1	Shirt collar	Around collar. Annotation goes down to approximately the first button in the frontal opening
2	Shirt sleeves	What is annotated are the *cuffs*. Annotation adjusted to boundaries of the cuff (leaving some small extra space to ensure all relevant area is inside)
3	T-shirt collar	Annotation drawn from the border to the slightly below the hemline of the collar (approximately double the space between the border and the hemline)
4	T-shirt sleeves	Similar criteria to those for the *T-shirt collar*. Annotations are drawn around the hemline of the sleeve (leaving approximately double space in the inner part)
5	Jeans hips	Jeans hips are annotated completely covering the belt loop (and a tiny bit more). If present, the pocket hole and the zip hemlines are included too
6	Jeans pants hemline	The area between the bottom of the pant and slightly above the hemline (approximately two thirds of space between the bottom of the pant and the hemline)
7	Polo collar	Around collar. Annotated down to approximately the first button in the frontal opening
8	Polo sleeves	Same criteria as for *T-shirt sleeves*
9	Sweater hood	Annotation starts at the beginning of the hood (no much extra space). If lace goes outside of the "hood area", it is ignored. Hood is annotated even if seen from the back
10	Sweater sleeves	Same criteria as for *Shirt sleeves*
11	Dress collar	The top part of the dress, including the holes for the arms

The dataset comprises 776 scenes of textile items belonging to six garment types: polo, jeans, t-shirt, dress, shirt and sweater. For each scene the dataset includes color image, point cloud, segmentation mask, and annotations. For each garment, one or two parts of the object have been manually annotated, with each class having between approx. 100 and 225 such annotations in the whole dataset. In Table 1 the different types of garments and annotated parts are described. This dataset is the extension of the one used in Ramisa et al. (2012), and we have made it publicly available for download.[1]

The data was acquired in a laboratory setting. The garments were laying on a gray table taking about one-third of the picture, and a Kinect camera was set up above at approximately 70 cm with a zenithal view (see Fig. 4). We used the default camera calibration matrices provided by the manufacturer.

The RGB-D images acquired with the camera have 640×480 pixels, and are provided in PNG format (image part), and in *PCD v.7* plain text file format (depth part). Some examples, with overlaid annotations, can be seen in Fig. 5.

The primary illumination source consists of fluorescents that provide a pure white light. Images include one or more garments. Because of the size and flexibility of the clothing objects, they can be partially out of frame, and occlusions may occur. Not all interest parts are always visible for every clothing item due to folds or occlusions. All classes are represented by at least two distinct object instances.

Fig. 4. Setup used to acquire the dataset. Garments lay on the table in different positions and deformations. RGB-D images are acquired using a Kinect camera looking downwards on a WAM robot cell.

Segmentation mask: For each image, a binary segmentation mask is provided. The mask is a 8-bit gray-level PNG image, with white pixels (value 255) belonging to the main garment, and black pixels (value 0) to the background. The segmentation masks have been generated via a combination of color and depth

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

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

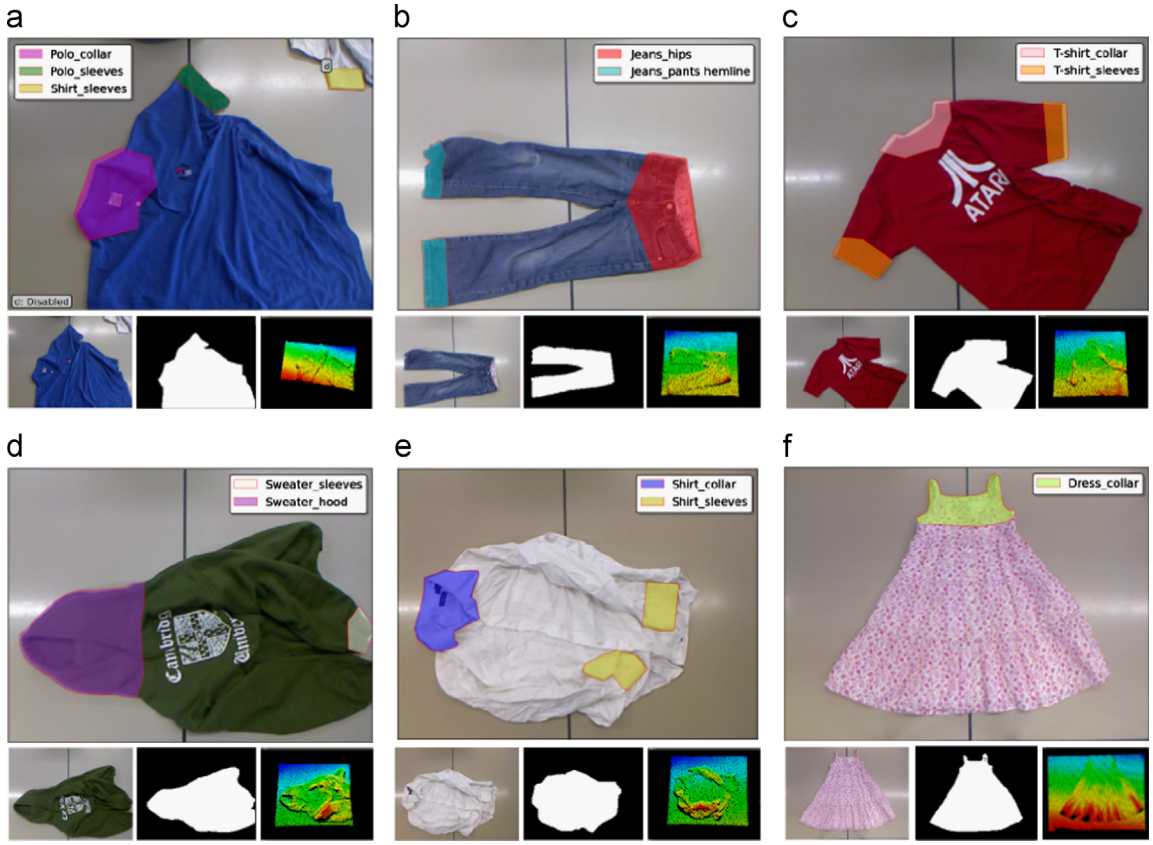

Fig. 5. Garment classes present in the dataset. For each of the six classes, a panel with four images is displayed: overlaid annotations (top), the original image (bottom left), the segmentation mask (bottom center) and a representation of the depth information provided by the Kinect camera (bottom right). (a) **Polo**: collar and sleeves, (b) **Jeans**: hips and hemline, (c) **T-shirt**: collar and sleeves, (d) **Sweater**: hood and sleeves, (e) **Shirt**: collar and sleeves, and (f) **Dress**: collar.

segmentation, and manually repaired in case of error of the automatic segmentation.

Annotation methodology: The ground truth data comes in the form of polygons tightly enclosing the clothing part. Since the method tested in this work uses bounding boxes, a rectangle tightly enclosing the annotated polygon is determined for each clothing part. The criteria used to determine what constitutes the clothing part can be seen in Table 1. The annotations are stored in a self-explanatory plain text ground truth file.

5. Experimental results

The objective of this work is to evaluate to what extent can a Bag of Visual Words based detection approach be used in the context of garment part detection for robot manipulation. With this objective, we have evaluated the part detection method described in Section 3 on the proposed dataset, using different combinations of descriptors.

Here is an overview of the experiments performed: first, we discuss the performance of the baseline method with the different descriptor combinations in Section 5.2. Then, in Section 5.3, we evaluate the performance of the Image-Level Prior classifiers, and its effects on the precision of the proposed method. Next, in Section 5.4, we also evaluate the proposed method in the CTU Spread Garments Dataset. Finally, in Section 5.5, we review the computational cost of the proposed method and of the different appearance and depth descriptors.

5.1. Experimental setup

In this section we describe the technical details of the methods and the settings used in the experiments.

Subsets: In order to assess how sensitive is the method to the degree of wrinkledness of the objects, we constructed two (partly overlapping) subsets: a baseline subset we called *Complete*, and the *Easy* subset where only mostly unoccluded and unwrinkled images of the parts are considered. In both subsets, approximately 30% of the examples of each class (between 15 and 65, depending on the class and the subset) are used for testing and the rest for training (between 40 and 155). Object instances are mixed in the training and testing sets (i.e. images of two shirts in the dataset can be found both in the training and testing sets). Both subsets consists only of images showing a single garment.

Performance measures: In our experiments, a part detection is considered a true positive if the center[2] of the detection bounding box falls within the ground truth area. To measure the performance, we use recall at k (R@K), which tells us how likely is the method to correctly detect the part if present in the image looking only at the k highest scored detections. This measure is relevant for robotic manipulation, since typically the robot arm will only be able to consider a few options in its planning, and the state of the garment will change after interaction. We also evaluate the results with the Average Precision (AP), commonly used in computer

[2] A substitute for a grasping point whose actual computation is outside the scope of this paper.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

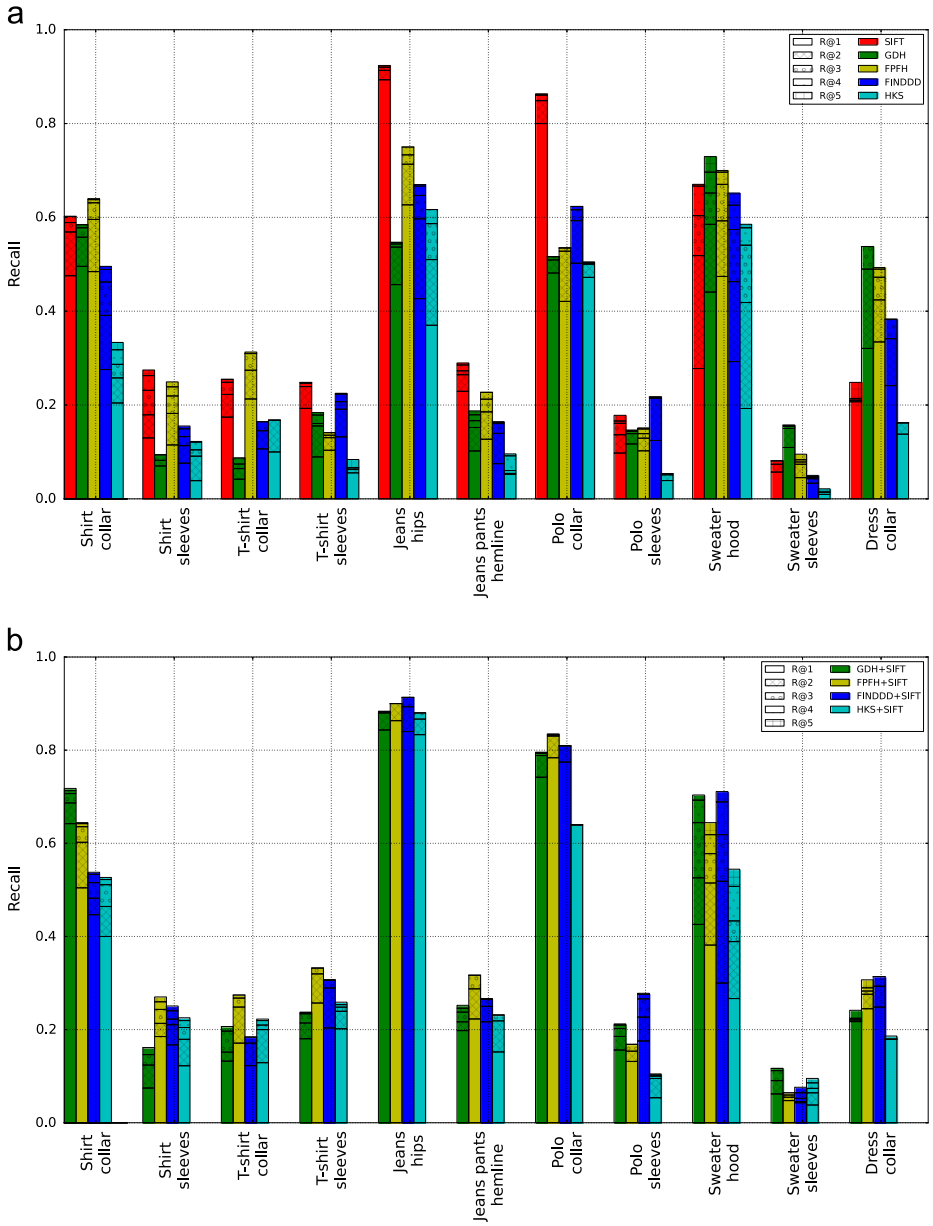

Fig. 6. Recall at *k* for each garment part category when using the different descriptors. The recall levels are stacked together (using different patterns) in a column of each descriptor (using colors) and garment part type. Plots correspond to (a) single descriptors and (b) combinations of SIFT and a shape descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

vision competitions, such as the Pascal Visual Object Classes Challenge.[3] The Average Precision corresponds to the area under the precision-recall curve, which shows the precision obtained at every different level of recall (i.e. how many false positives were encountered before finding each one of the positives in the test set). The AP allows to express the performance of a method with a single number, losing, however, particular information on its behavior at different recall levels. The mean Average Precision (mAP) is the mean of the AP scores obtained across all classes. All results presented are the average of 10 repetitions of the full training procedure, with fixed training and test sets to make results comparable.

Number of visual words: The optimal number of visual words in the dictionary is a parameter highly dependent on the environment in which the detector will work. Each visual word should ideally be activated with the image of a particular physical element while masking the internal variability of the element type. Automatically determining the correct size for a vocabulary is an active research topic, and sophisticated dictionary learning techniques have been proposed to address this problem (Winn et al., 2005; Fulkerson et al., 2008; Mairal et al., 2008). After evaluating with a small number of vocabulary sizes in a subset of the data, we found that 128 visual words is a good choice for our experiments, as it had a competitive performance with all descriptors. This number may seem significantly smaller than the one typically used in object recognition benchmarks, but it has to be noticed that the "visual world" encountered in this dataset is much smaller than that of, for example, the Pascal VOC dataset, and hence a smaller number of

[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

visual words may suffice to describe it. Nevertheless, using an appropriate dictionary learning technique could improve the results, and is left for future work.

Descriptor and detector parameters: Default descriptor parameters were used when possible, and sampling was done densely every 6 pixels for all descriptors (including SIFT), as it is a good trade-off between computational power and descriptor density. Based on a small-scale cross-validation study, we determined the following parameters for the depth descriptors: we set the patch size for HKS to 60 × 60 pixels, and keep the first 60 frequencies as the final descriptor; we disabled the illumination coefficients for the GDH, and set 8 bins for geodesic distance and 11 for depth, resulting in a 88-dimensional vector. For the FINDDD descriptor, the bin centers are generated taking the vertexes of the triangles obtained using a sphere tessellation algorithm applied only to the north hemisphere, which in our case yielded 13 and 41 vertexes at the two lowest tessellation resolution levels and, based on empirical results, we selected the former as it adapted better to the level of noise of the Kinect sensor. The patch size for FINDDD is 43 pixels, and the number of spatial subdivisions is 16. FPFH has been used with the default parameters.

Regarding the parameters of the detection method, we adjusted the size and shape of the sliding windows based on those of the ground truth annotations of the training set. Classifier parameters were determined using cross-validation on the training set.

5.2. Results for the detection of garment parts

The results of applying the descriptors to the *Complete* subset show a large variability between classes, which is nevertheless very consistent for all descriptors (Fig. 6a). Classes that correspond to large, distinctive and less deformable parts of the garments, like the shirt or polo collar, or the jeans hips, had a much better performance, notably with the SIFT appearance descriptor. On the other hand, classes corresponding to smaller, less distinctive parts led to a bad performance for both appearance and shape descriptors alike. It is also noteworthy the improvement in performance attained by the 3D descriptors in some of the classes, like the dress collar. Regarding the combination of SIFT and a 3D descriptor (Fig. 6b), the performance is similar or better to the stand-alone descriptors, except for some classes like the dress collar, where the SIFT descriptor has a detrimental effect.

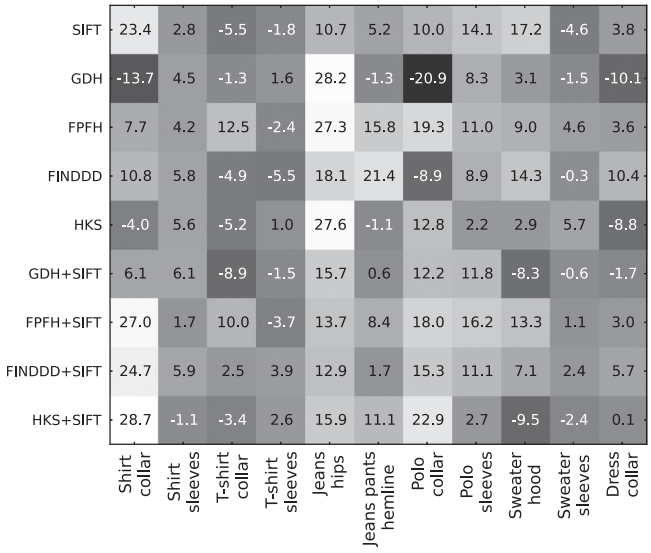

Fig. 7. Relative improvement of recall at one between the *complete* subset and the *easy* subset. Results improve but not significantly, suggesting that the problem remains difficult.

Regarding the *easy* subset, the relative improvement in recall at one w.r.t the *Complete* subset is presented in Fig. 7. Overall, the results improve a bit, but there are some cases where the contrary occurs, thus it seems that the problem remains difficult, and the difference on performance may be more related to the particular train/test splits. The average improvement is 6%, which suggests that the proposed method has capacity to handle complex scenes. Table 2 displays the average differences across the classes for each descriptor combination. Jeans and shirts are the garment types that show more improvement with the reduced complexity. It is also noticeable that single descriptors have a more erratic behavior (notably GDH) than combinations, which are able to obtain more leverage from the easier dataset.

Table 2
Average variation in recall at one in the *easy* subset with respect to the *complete* subset. Descriptor names with the + symbol denote combinations with SIFT.

SIFT	GDH	FPFH	FINDDD	HKS	GDH+	FPFH+	FINDDD+	HKS+
6.8	−0.3	10.3	6.4	3.5	2.8	9.9	8.5	6.1

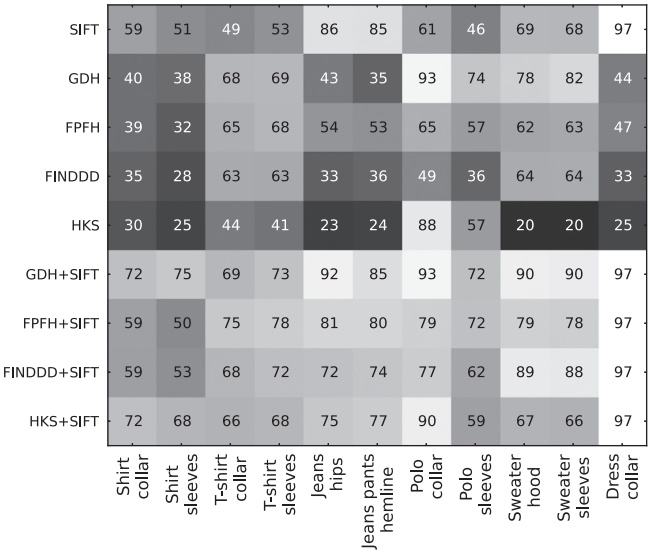

Fig. 8. Average Precision classification performance of the Image-Level Prior classifiers.

Table 3
Classification results of the Image-Level Prior classifiers. C stands for *complete*, E for *easy* subsets; mAP stands for mean Average Precision, and mACC for mean Accuracy.

Descriptor	C		E	
	mAP	mACC	mAP	mACC
SIFT	65.8	80.1	55.5	78.4
GDH	60.3	78.5	47.4	78.4
FPFH	55.0	76.3	43.6	76.4
FINDDD	45.8	73.4	38.6	73.0
HKS	35.9	64.3	31.7	63.7
GDH+SIFT	**82.4**	**89.2**	**70.5**	**88.3**
FPFH+SIFT	75.3	85.8	61.4	82.5
FINDDD+SIFT	73.6	82.4	60.5	81.5
HKS+SIFT	73.2	83.1	60.4	81.2

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

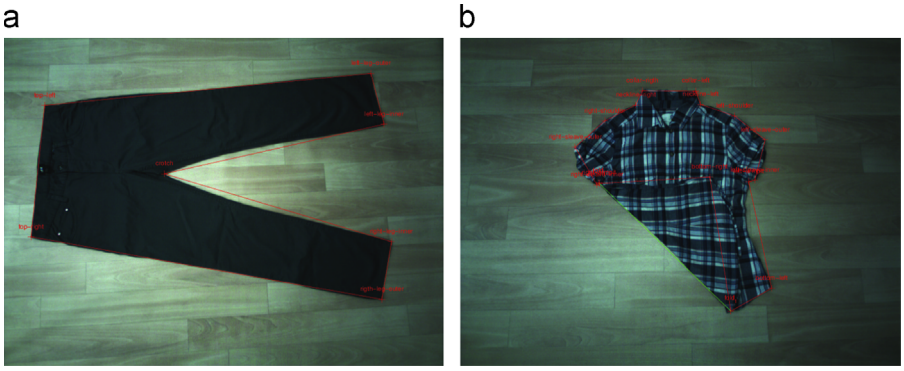

Fig. 9. Example images and annotations of the CTU Spread Garments Dataset. Best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 4
Mean Average Precision of the proposed method with and without the Image-Level Prior. C stands for *Complete* and E for *Easy* subsets.

Descriptor	Mean Average Precision			
	C	C+ILP	E	E+ILP
SIFT	17.3	18.7	22.8	22.9
GDH	6.8	7.7	4.0	5.4
FPFH	6.3	7.8	10.7	11.9
FINDDD	5.8	6.4	6.5	7.2
HKS	2.8	3.7	2.9	3.9
GDH+SIFT	18.7	20.4	20.5	22.5
FPFH+SIFT	21.9	**24.0**	**29.2**	28.3
FINDDD+SIFT	20.1	20.5	23.5	24.7
HKS+SIFT	15.1	15.7	20.5	20.8

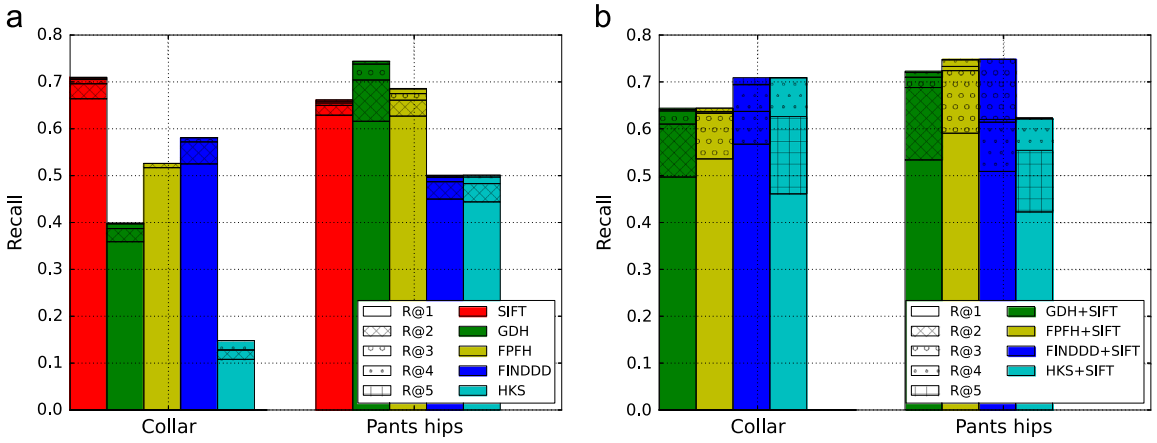

Fig. 10. Recall at *k* for the *Collar* and *Pant hips* categories in the CTU Spread Garments dataset when using the different descriptors. Same layout as in Fig. 6.

5.3. Precision of the detector

The previous results reflected how likely the method is to correctly select the relevant part if it is present in the image; however, for robots working in an unstructured environment, it is also important to take into consideration precision.[4] As previously said, in order to take into account both precision and recall, we have used the Average Precision (AP) measure obtained when running the method in each image of the test set, even those that did not contain the part of interest.

In order to reduce the number of false positives, we evaluate the use of an Image Level Prior (ILP) to discard images not likely to contain the part of interest. A logistic regression classifier is trained for each part, and its score is used to select which images are searched for the part of interest and which are discarded.

Classification results: Fig. 8 shows the Average Precision of the classifiers used as ILP for each class individually in the *Complete* subset. The combinations of appearance and depth descriptors achieve consistently about 20 percentage points more than the descriptors alone for this task. Table 3 shows the mean AP and mean accuracy results for each subset and descriptor combination. As can be seen, GDH+SIFT attains the highest score across both datasets and evaluation measures, followed by FPFH+SIFT. These

[4] Precision = TP/(TP+FP), where TP are the true positives and FP the false positives. Recall = TP/(TP+FN), where FN are the false negatives.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

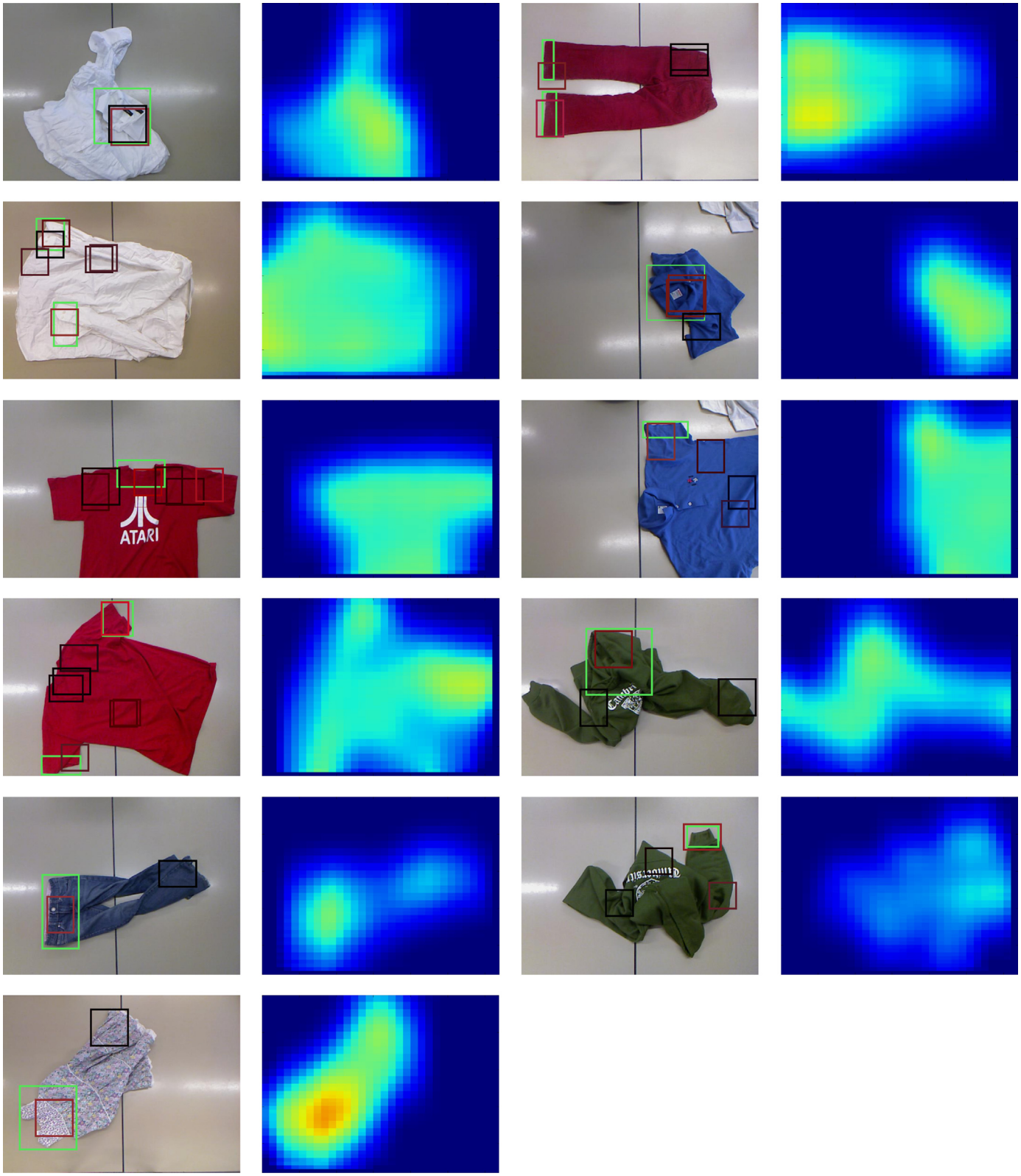

Fig. 11. Example of detection results, after applying the χ^2 RBF Support Vector Machine, and probability maps, generated with the logistic regression classifier, for each part in the dataset. The green bounding boxes correspond to the annotated ground truth, and predicted detections are shown as a bounding box with color from red to black according to its score normalized by the maximum score of the detections in the image. The results correspond to the combination of SIFT and FPFH. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

performances seem good enough to reliably select which classes to search in an image without significantly affecting the recall.

Detection results with ILP: Table 4 shows the mean detection AP (mAP) across all classes obtained with and without the ILP, on the two considered subsets. Using the ILP the precision of the results improves on average, although in some cases they degrade as correct detections are discarded when the ILP fails. The best mAP is consistently obtained by the combination of FPFH and SIFT

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

Table 5
Mean Average Precision of the proposed method only considering the classes *shirt collar*, *jeans hips* and *polo collar*.

Descriptor	Mean Average Precision			
	C	C+ILP	E	E+ILP
SIFT	48.7	48.3	66.0	58.2
GDH	12.4	12.6	6.8	8.7
FPFH	10.8	12.0	19.7	16.9
FINDDD	12.2	11.5	8.2	9.5
HKS	7.7	8.9	8.6	9.7
GDH+SIFT	52.3	53.4	55.2	57.5
FPFH+SIFT	**53.6**	52.1	**68.9**	60.7
FINDDD+SIFT	49.3	47.1	57.3	58.0
HKS+SIFT	43.4	41.3	62.8	58.1

Table 6
Mean Average Precision with and without ILP of the proposed method in the CTU Spread Garments dataset considering the classes **collar** and **pants hips**. See text for details on how the training and testing sets are constructed.

Descriptor	Collar		Pants hips	
	CTU	CTU+ILP	CTU	CTU+ILP
SIFT	50.6	55.4	49.4	52.4
GDH	30.5	29.9	42.8	38.7
FPFH	41.0	36.9	46.8	42.6
FINDDD	35.2	33.8	32.2	26.3
HKS	14.2	11.8	27.0	20.7
GDH+SIFT	51.5	52.1	53.4	52.9
FPFH+SIFT	56.3	**57.4**	59.1	**59.2**
FINDDD+SIFT	53.8	56.3	50.9	52.1
HKS+SIFT	44.2	48.7	42.3	43.8

across all subsets. These results also show that, although not manifestly increasing recall, the addition of 3D descriptors helps increase the precision of the method (Fig. 10). Fig. 11 shows qualitative examples of the method performance.

Since the proposed method seems unsuitable in its current form for some of the clothing parts in the dataset (judging from the results in Table 4), we also compute the aggregated results focusing on the classes that offer a performance acceptable for its direct usage in robotic manipulation experiments (i.e. shirt collar, jeans hips and polo collar). These results are shown in Table 5. As can be seen, combining SIFT with the depth based descriptors often improves the results, suggesting that depth information helps the method generalize better. In terms of recall at one, average results over 70% are consistently obtained for these classes.

5.4. Detection results in the CTU spread garments dataset

In order to further test our method we selected the CTU Color and Depth Image Dataset of Spread Garments (Wagner et al., 2013), since it includes appearance and depth data, as well as annotations.[5] Fig. 9 shows two example images of this dataset with the corresponding annotations.

We have selected all the available images with associated depth data depicting a collar (which includes shirts, polo shirts and coats), and a similar number of images with trousers (more abundant in the dataset). This amounted to 600 images, from which we separated one third for testing and the rest for training. For each image, several fixed annotated points that yield a rough outline of the garment are provided. From these points we have automatically derived bounding box annotations compatible with our method, as well as segmentation masks, not very accurate sometimes, but sufficient for our experiments. Table 6 shows the results obtained in this dataset. As can be seen, results for the evaluated classes are similar to those obtained in our proposed dataset. Performance of combinations is, in general, better than descriptors alone (also for SIFT), and the combination of FPFH and SIFT attains the best results, with the combination of FINDDD and SIFT following closely.

5.5. Computational cost

In terms of computational cost, we compare the wall-clock time of the different descriptors, taking into account the implementation differences. The experiments have been performed on a 3.33 GHz Linux machine.

SIFT,[6] FPFH and FINDDD[7] are implemented in C++, and therefore their computation times are directly comparable. SIFT and FINDDD obtained times around 1 s per RGB-D scan, the former being slightly faster, and FPFH took around 322 s. For the GDH and the HKS, we use respectively a modification of the original GIH code[8] and an in-house implementation. In both cases, the implementations are in Matlab, with the most time-consuming parts written in C. Consequently, the time taken for these descriptors is not directly comparable to that of the descriptors implemented in C++. However, we make some observations on its performance that lead us to think that they would still be slower if completely implemented in C++: the wall-clock time for a RGB-D scan using the GDH was 6826 s, and 77% of this time was spent computing the geodesic curves, using the low-level `contourc` Matlab routine, and the HKS took 20,703 s on average, and 40% of the time was spent finding the eigenvalues of the Laplacian using the `eigs` Matlab routine.

Regarding the computational cost of the complete method, it typically takes around 1 or 2 s per image using a single core (excluding descriptor computation) using the efficient sliding window approach of Wei and Tao (2010).

6. Conclusion

In this work, we have introduced a benchmark for the problem of perception of clothes for manipulation purposes, and we performed a comparative evaluation of four 3D descriptors, alone and combined with the SIFT appearance descriptor.

In general, results show that single-shot detection of clothing parts is a difficult task, but that it is possible to attain a reasonable performance for certain characteristic garment parts, like jeans hips or shirt and polo collars.

From the evaluated 3D descriptors, FPFH obtained an overall higher performance in terms of Average Precision and Recall at one. However, the computational cost of this descriptor makes FINDDD – which attained comparable results, specially when combined with SIFT – an interesting alternative for near real time applications.

When compared with the SIFT appearance descriptor, the performance of the 3D descriptors in terms of recall is slightly lower; however, the combination of the 3D descriptors with SIFT maintained or slightly improved the recall, and had a positive

[5] http://clopema.felk.cvut.cz/color_and_depth_dataset.html

[6] We used the C++ implementation from the *VLFeat* library http://www.vlfeat.org/

[7] FPFH and FINDDD are implemented in C++ using PCL www.pointclouds.org

[8] http://www.dabi.temple.edu/~hbling/code_data.htm

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

effect in the AP, suggesting that this 3D information would help generalize better to previously unseen clothes.

In order to increase the precision, we evaluated an Image-Level Prior (ILP) to directly discard images not likely to contain the part of interest. The results show an overall improvement in AP, but in some cases classification errors cause a decrease in recall, which in turn impacts the AP score. Regarding the classification performance, the combination of a 3D descriptor and SIFT significantly outperforms any of the descriptors alone.

Another contribution of this work is a novel dataset of RGB-D scans of garments laying on a flat surface. Specific parts of the garments have been manually annotated with polygons; and a segmentation mask, which selects the textile object, is provided for each scan. The dataset is aimed at evaluating part detection, classification and segmentation methods for textile objects under severe deformations. To our knowledge, this is the first dataset of this kind, and we hope it encourages progress in perception methods for highly deformable textile objects.

One common characteristic of the 3D descriptors evaluated in this work is their high sparsity. A dimensionality reduction technique (like principal component analysis, with the assumption that the data is normally distributed) could help decorrelate the components and remove the noise.

Finally, we would like to leave the core dataset presented open to extensions, e.g. incorporating more instances of the different garments to allow for a better testing of the generalization properties of the descriptors.

Acknowledgments

This research is partially funded by the Spanish Ministry of Science and Innovation under Project PAU+ DPI2011-2751, the EU Project IntellAct FP7-ICT2009-6-269959 and the ERA-Net Chistera Project ViSen PCIN-2013-047. A. Ramisa worked under the JAE-Doc grant from CSIC and FSE.

References

Aldavert, D., Lopez de Mantaras, R., Ramisa, A., Toledo, R., 2010. Fast and robust object segmentation with the Integral Linear Classifier. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, San Francisco, CA, USA, pp. 1046–1053. http://dx.doi.org/10.1109/CVPR.2010.5540098.

Aragon-Camarasa, G., Oehler, B., Liu, Y., Sun, L., Cockshott, P., Siebert, J.P., 2013. Glasgow's Stereo Image Database of Garments. Technical Report. CoRR abs/1311.7295.

Bronstein, M., Kokkinos, I., 2010. Scale-invariant heat kernel signatures for non-rigid shape recognition. In: Computer Vision and Pattern Recognition (CVPR), pp. 1704–1711.

Csurka, G., Dance, C.R., Fan, L., Bray, C., Willamowski, J., 2004. Visual Categorization with Bags of Keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–22.

Cusumano-Towner, M., Singh, A., Miller, S., O'Brien, J.F., Abbeel, P., 2011. Bringing clothing into desired configurations with limited perception. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA11), Shangai, China. pp. 3893–3900.

Doumanoglou, A., Kargakos, A., Kim, T., Malassiotis, S., 2014. Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA14), Hong Kong, China, pp. 987–993.

Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874, http://dx.doi.org/10.1145/1390681.1442794.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vis. 59, 167–181, http://dx.doi.org/10.1023/B:VISI.0000022288.19776.77.

Fulkerson B., Vedaldi A., Soatto S., Localizing objects with smart dictionaries In: European Conference on Computer Vision (ECCV), 2008, Springer, Marseille, France, 179–192.

Grady, L., 2006. Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1768–1783, http://dx.doi.org/10.1109/TPAMI.2006.233.

Hidayati, S., Cheng, W., Hua, K., 2012. Clothing genre classification by exploiting the style elements. In: Proceedings of the 20th ACM International Conference on Multimedia, ACM, Nara, Japan, pp. 1137–1140. http://dx.doi.org/10.1145/2393347.2396402.

Himmelsbach, M., Luettel, T., Wuensche, H.J., 2009. Real-time object classification in 3D point clouds using point feature histograms. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, St. Louis, USA, pp. 994–1000. http://dx.doi.org/10.1109/IROS.2009.5354493.

Janoch, A., Karayev, S., Barron, J.T., Fritz, M., Saenko, K., Darrell, T., 2011. A category-level 3-D object dataset: putting the kinect to work. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, Barcelona, Spain, pp. 1168–1174. http://dx.doi.org/10.1109/ICCVW.2011.6130382.

Lai, K., Bo, L., Ren, X., Fox, D., 2011. A large-scale hierarchical multi-view RGB-D object dataset. In: 2011 IEEE International Conference on Robotics and Automation, IEEE, Shangai, China, pp. 1817–1824. http://dx.doi.org/10.1109/ICRA.2011.5980382.

Lázaro-Gredilla, M., Gómez-Verdejo, V., Parrado-Hernández, E., 2012. Low-cost model selection for SVMs using local features. Eng. Appl. Artif. Intell. 25, 1203–1211, http://dx.doi.org/10.1016/j.engappai.2012.05.021.

Li, Y., Olson, E.B., 2010. Extracting general-purpose features from LIDAR data. In: 2010 IEEE International Conference on Robotics and Automation, IEEE, Anchorage, Alaska, USA, pp. 1388–1393. http://dx.doi.org/10.1109/ROBOT.2010.5509690.

Lin, H., Lin, C., Weng, R., 2007. A note on Platt's probabilistic outputs for support vector machines. Mach. Learn. 68, 267–276.

Ling, H., Jacobs, D.W., Deformation invariant image matching. In: IEEE International Conference on Computer Vision, Beijing, China, 2005, pp. 1466–1473.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A., 2008. Supervised Dictionary Learning. Technical Report. INRIA.

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., Abbeel, P., 2010. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In: Proceedings of the IEEE international Conference on Robotics and Automation (ICRA10), pp. 2308–2315.

Mariolis, I., Malassiotis, S., 2013. Matching folded garments to unfolded templates using robust shape analysis techniques. Computer Analysis of Images and Patterns. Lecture Notes on Computer Science 8048, 193–200, http://dx.doi.org/10.1007/978-3-642-40246-3_24.

Moreno-Noguer, F., Salzmann, M., Lepetit, P., Fua, P., 2009. Capturing 3D stretchable surfaces from single images in closed form. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1842–1849.

Moreno-Noguer, F., 2011. Deformation and illumination invariant feature point descriptor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1593–1600.

Moreno-Noguer, F., Fua, P., 2011. Stochastic exploration of ambiguities for non-rigid shape recovery. IEEE Trans. Pattern Anal. Mach. Intell. 35, 463–475.

Parkhi, O., Vedaldi, A., Jawahar, C., Zisserman, A., 2011. The truth about cats and dogs. In: International Conference on Computer Vision, pp. 1427–1434.

Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74.

Ramisa, A., Alenyà, G., Moreno-Noguer, F., Torras, C., 2012. Using depth and appearance features for informed robot grasping of highly wrinkled clothes. In: 2012 IEEE International Conference on Robotics and Automation, IEEE, St. Paul, MN, USA, pp. 1703–1708. http://dx.doi.org/10.1109/ICRA.2012.6225045.

Ramisa, A., Alenyà, G., Moreno-Noguer, F., Torras, C., 2013. FINDDD: a fast 3D descriptor to characterize textiles for robot manipulation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Tokyo, Japan, pp. 824–830. http://dx.doi.org/10.1109/IROS.2013.6696446.

Rashedi, E., Nezamabadi-pour, H., 2013. A stochastic gravitational approach to feature based color image segmentation. Eng. Appl. Artif. Intell. 26, 1322–1332, http://dx.doi.org/10.1016/j.engappai.2012.10.002.

Rusu, R., Marton, Z., Blodow, N., Beetz, M., 2008. Persistent point feature histograms for 3D point clouds. In: Intelligent Autonomous Systems 10: IAS-10, p. 119.

Rusu, R.B., Blodow, N., Beetz, M., 2009. Fast Point Feature Histograms (FPFH) for 3D registration. In: 2009 IEEE International Conference on Robotics and Automation, IEEE, Kobe, Japan, pp. 3212–3217. http://dx.doi.org/10.1109/ROBOT.2009.5152473.

Sanchez, J., Ostlund, J., Fua, P., Moreno-Noguer, F., 2010. Simultaneous pose, correspondence and non-rigid shape. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1189–1196.

Sethian, J., 1996. A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. 93, 1591–1595.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Real-time human pose recognition in parts from single depth images. In: Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1297–1304.

Shotton, J., Johnson, M., Cipolla, R., Center, T., Kawasaki, J., Semantic texton forests for image categorization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 2008, pp. 1–8. http://dx.doi.org/10.1109/CVPR.2008.4587503.

Sun, J., Ovsjanikov, M., Guibas, L., 2009. A concise and probably informative multi-scale signature based on heat diffusion. In: Computer Graphics Forum, pp. 1383–1392.

Tangelder, J., Veltkamp, R., 2004. A survey of content based 3D shape retrieval methods. In: Proceedings Shape Modeling Applications, IEEE, Genova, Italy, pp. 145–388. http://dx.doi.org/10.1109/SMI.2004.1314502.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025

Wagner, L., Krejčová, D., Smutný, V., 2013. CTU Color and Depth Image Dataset of Spread Garments. Technical Report. CTU-CMP-2013-25.
Wang, P., Miller, S., Fritz, M., Darrell, T. Perception for the manipulation of socks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2011, pp. 4877–4884.
Wei, Y., Tao, L., 2010. Efficient histogram-based sliding window. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3003–3010, http: //dx.doi.org/10.1109/CVPR.2010.5540049.
Willimon, B., Birchfield, S., Walker, I., 2011. Classification of clothing using interactive perception. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA11), pp. 1862–1868.
Willimon, B., Walker, I., Birchfield, S., 2013. Classification of clothing using midlevel layers. ISRN Robot. 2013, 1–17, http://dx.doi.org/10.5402/2013/630579.
Winn, J., Criminisi, A., Minka, T. Object categorization by learned universal visual dictionary. In: Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 2005, pp. 1800–1807. http://dx.doi.org/10.1109/ICCV.2005.171.
Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L., 2012. Parsing clothing in fashion photographs. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, Providence, Rhode Island, USA, pp. 3570–3577. http://dx.doi.org/10.1109/CVPR.2012.6248101.
Yamazaki, K., Inaba, M., 2013. Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Tokyo, Japan, pp. 2710–2717. 978-1-4673-6357-0/13.
Yang, H.Y., Wang, X.Y., Zhang, X.Y., Bu, J., 2012. Color texture segmentation based on image pixel classification. Eng. Appl. Artif. Intell. 25, 1656–1669, http://dx.doi.org/10.1016/j.engappai.2012.09.010.
Zhang, J., Marszaek, M., Lazebnik, S., Schmid, C., Marszalek, M., 2006. Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vis. 73, 213–238, http://dx.doi.org/10.1007/s11263-006-9794-4.

Please cite this article as: Ramisa, A., et al., Learning RGB-D descriptors of garment parts for informed robot grasping. Eng. Appl. Artif. Intel. (2014), http://dx.doi.org/10.1016/j.engappai.2014.06.025