
Article:
Martínez, L., Ruiz-del-Solar, J., Sun, L. et al. (2019) Continuous perception for deformable objects understanding. Robotics and Autonomous Systems, 118. pp. 220-230. ISSN: 0921-8890

https://doi.org/10.1016/j.robot.2019.05.010

Continuous Perception for Deformable Objects Understanding

Luz Martínez[a,b], Javier Ruiz-del-Solar[a], Li Sun[b], J. Paul Siebert[b], Gerardo Aragon-Camarasa[b]

[a]*Advanced Mining Technology Center & Dept. of Elect. Eng., Universidad de Chile.*
[b]*School of Computing Science, University of Glasgow, Glasgow, UK.*

Abstract

We present a robot vision approach to deformable object classification, with direct application to autonomous service robots. Our approach is based on the assumption that continuous perception provides robots with greater visual competence for deformable objects interpretation and classification. Our approach thus classifies the category of clothing items by continuously perceiving the dynamic interactions of the garment's material and shape as it is being picked up. Our proposed solution consists of extracting continuously visual features of a RGB-D video sequence and fusing features by means of the Locality Constrained Group Sparse Representation (LGSR) algorithm. To evaluate the performance of our approach, we created a fully annotated database featuring 150 garment videos in random configurations. Experiments demonstrate that by continuously observing an object deform, our approach achieves a classification score of 66.7%, outperforming state-of-the-art approaches by a $\sim$ 27.3% increase.

Keywords: Deformable Object Classification, Continuous Perception, Robot Vision

1. Introduction

Autonomous recognition and handling of deformable objects is an essential and challenging task for autonomous service robots. In this paper, we state that a continuous perception approach equips a robot to recognize deformable objects from a random configuration as the robot picks them up from a flat surface. Deformable objects comprise clothing, linens and produce, to name a few; and, we focus on clothing in this paper since it can take practically an infinite range of possible configurations, ranging from a relatively smooth state to a crumpled state.

Perceiving actions and states of objects in the environment should become a standard requirement for robots to be deployed in domestic environments and service scenarios such as hotels and hospitals to mitigate failures and accidents. We, humans, have exceptional capabilities to manipulate and interact with deformable objects. The reason is that our vision system senses the environment continuously, accumulates predictions and creates relations over time about the state of objects, people and the environment, including highly-deformable objects. Hence, the key is to observe the state of the object continuously but current approaches for deformable object visual perception focuses on recognizing or classifying the state of an object from one frame, then plan the most optimal action, and, finally, execute the action. State-of-the-art approaches have indeed solved complex tasks such as pick-and-place tasks [1] and clothing perception and manipulation [2][3][4] but none have investigated if continuous perception increases classification and recognition rates of deformable objects.

Hence, we describe an approach for deformable object classification based on continuously perceiving the object's state from the moment it is picked up from a working table. The target robotic tasks are pick-and-place, garment sorting and folding and unfolding scenarios, to name a few. The underlying idea is to extract visual features from 2.5D images in consecutive frames to learn a temporal-consistent representation of the clothing's dynamic attributes. For this, a deformable object is placed in a random-configuration on a flat surface where the robot grasps it and starts observing the object's physical deformation. To pick up the objects, we employ a basic, yet powerful heuristic grounded on the highest observable point using depth information. Once the robot grasps the object, the robot goes to a pre-

Email addresses: luz.martinez@amtc.cl (Luz Martínez), jruizd@ing.uchile.cl (Javier Ruiz-del-Solar), paul.siebert@glasgow.ac.uk (J. Paul Siebert), gerardo.aragoncamarasa@glasgow.ac.uk (Gerardo Aragon-Camarasa)

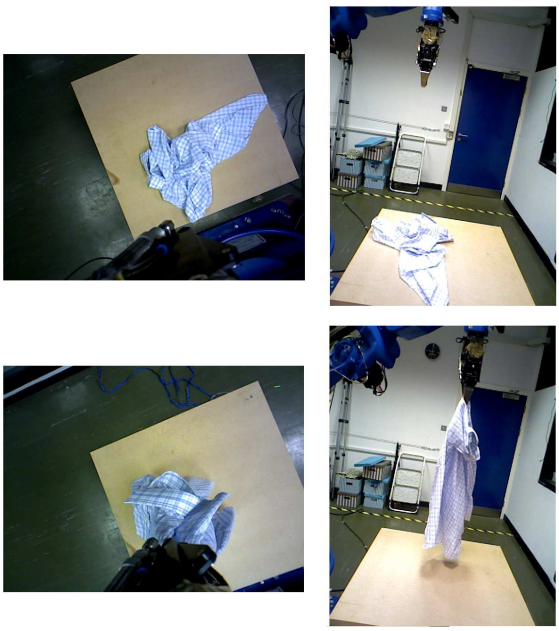

Figure 1: First and final RGB images of the video sequences from our continuous perception database from the top and side views.

defined position above the table while capturing frames from a top and side views, i.e., egocentric and exocentric views. The initial and final images of a typical image sequence can be observed in Figure 1.

The key contributions of this paper are:

1. We present and demonstrate a continuous visual perception approach for deformable object classification while a robot picks and observes how a deformable object changes over time.
2. We have conducted an extensive ablation study to investigate how visual features approaches contribute and perform to the classification of deformable objects under continuous perception.
3. We describe a database of different clothing items as they are being picked up by a robot, which we use to validate our continuous perception approach alongside with state-of-the-art clothing databases[5].

The visual feature framework approach we adopted is inspired by, and builds on [6] and [3]. In this paper, we expand this framework by integrating continuous visual knowledge as being extracted from a video sequence; demonstrating, for the first time, a functional continuous visual perception approach to deformable objects understanding. Similarly, Our database is the first fully annotated collection of video sequences in the literature and can be used for further comparison and benchmarking for continuous and single-frame classification and recognition [7]. Our database can be downloaded from http://dx.doi.org/10.5525/gla.researchdata.669.

This paper is organized as follows: Section 2 presents background research in robot vision for classifying and recognizing deformable objects. Section 3 describes the continuous perception approach to deformable object understanding while Section 4, our finding. Finally, discussion and conclusions are given in Section 5.

2. Related Work

Current approaches for deformable object recognition and classification can be divided into two categories; those that recognize a deformable object when it is on a table [8] [9] [6], and those that recognize deformable objects when they are hanging from a robot's gripper [10] [11]. In this paper, we merge both categories by using the sequence that starts in the first category and ends in the second.

When deformable object are on a table, approaches consists of classifying them from only one image (single-shot perception [6]) or that change the their configuration to increase the prediction reliability, relying on the randomness of the deformable object after interaction (interactive perception [3]). That is, Li et al. [6] showed that it is possible to classify deformable objects in unconstrained and random configurations from single image frames. They proposed to extract visual features to represent material physical attributes of garments from depth images. In their later work [3], the authors extended their system to interactively perceive clothing items by capturing image frames after the robot interacted with a garment to change its physical configuration. The latter approach demonstrated substantial improvement over single-frame approaches, which improved the classification confidence by increasing the number of observations after interaction. In this paper, we build on both approaches by allowing the robot to observe and understand how an object deforms as being picked up from a table. By employing a Locality Constrained Group Sparse Representation (LGSR) technique, our continuous perception approach encodes and creates temporal concepts of the object's physical dynamics for its classification. When the object hangs from a robot's gripper, it is common to take advantage of the classification for pose recognition and detect the optimal grasping points, which the robot can then plan subsequent actions such as unfolding for garments. These systems are devised as a two-stage process [12] [13]; where classification informs and reduces

the search space for 3D pose estimation for grasp planning.

Early research in deformable object recognition comprised extracting visual features using silhouette features [14] [15] [16]. Then, with the arrival of low-cost RGB-D cameras, approaches exploiting depth information were used. Most of these approaches match patches based on 3D local features such as *Geodesic-Depth Histograms* (GDH) [8], *Fast Point Feature Histograms* (FPFH) [9] [17], *Heat Kernel Signatures* (HKS) [18] and *Fast Integral Normal 3D* (FINDDD) [19]. Other approaches integrate a full 3D model from depth images to extract 3D volumetric features [20]. The common ground in previous research is that the most distinctive visual features are wrinkles, which indeed provide relevant information of the type of material as demonstrated in [8] [6] [3] with direct robotic applications to dual-arm flattening [21] [22], and ironing [23]. Moreover, the "wrinkledness" measure has been widely used in state-of-the-art algorithms. This measure uses entropy to analyze how much of the surrounding area of a point has normals aligned in the same orientation, i.e., a flat surface or a combination of few flat surfaces [24]. Advanced analysis of wrinkles has also been carried out, aiming to identify their shape and topology using a size based classification procedure, which requires detecting the length, width, and height of each wrinkle [22]. In this work, we adopt different visual features to investigate how they contribute to the performance of classifying deformable objects using continuous perception.

Recently, approaches based on Deep Neural Networks (DNNs) [25] [26] [27] have been used for deformable object recognition and classification. Although most of the systems use real images for training, others use simulated models to increase the amount of training data [12] [25] [26] [27]. However, deformable objects in a random configuration are highly challenging to simulate, and investigating continuous perception approaches for clothing classification is intractable at the moment. Moreover, the DNNs can achieve competitive classification performance through an end-to-end traning but lack of an interpretable analysis. In this paper, we leverage Locallity-constrained Linear Coding (LLC) [40, 6] and Gaussian Process Latent Variable Model (GPLVM) [41] for the model selection of the high-dimensional feature representation.

3. The Continuous Perception Approach

We claim that continuous perception equips autonomous systems with needed visual capabilities to classify and recognize the type of deformable object based on their physical attributes and distinctive visual characteristics. For this, we therefore investigated and selected the optimal combination of visual feature techniques to support continuous robotic perception. We also investigated how distinctive visual features of the objects contribute to the visual classification task at hand.

3.1. Experimental Techniques and Methods

The experimental setup consists of capturing multiple depth images from two different camera positions, namely, egocentric and exocentric views, while a robot arm grasps a garment from a flat surface. Depth images from both cameras are then passed to the continuous perception framework shown in Figure 2. This framework, inspired by [6], consists of 3 modules: (1) visual feature extraction and coding of local features, (3) integration of the features, and (4) temporal, continuous, classification.

In Figure 2, local visual features characterize unique information about the dynamics of deformable materials, while global features capture the overall shape of the object as it is manipulated. To maintain the focus on the ability of the robot to perceive and classify deformable objects continuously, we assume that objects are easily segmented from the working table. That is, we segmented items based on a simple height threshold algorithm with respect to the table, and this has been recorded in our experimental dataset [7]. For advanced segmentation algorithms, we refer the reader to [28] [29] [30].

In the first module of the continuous perception experimental setup, we selected local and global features that are the de-facto visual features to deformable object classification and, consequently, have shown good performance in state-of-the-art approaches for single frame classification [18, 24, 6, 3]. Local visual features are encoded and then concatenated to global features to create a condensed visual description for a given frame. We called this the Composite Feature Vector (CFV), Figure 3.4.1. CFV thus captures both the dynamic interactions plus the global shape of the object, which is then put into a temporal representation of the depth video sequence using the Locality-Constrained Group Sparse representation (LGRS) algorithm [31]. For completeness, we describe briefly the techniques and methods we adopted and implemented in the following sections to support continuous perception for deformable object understanding.

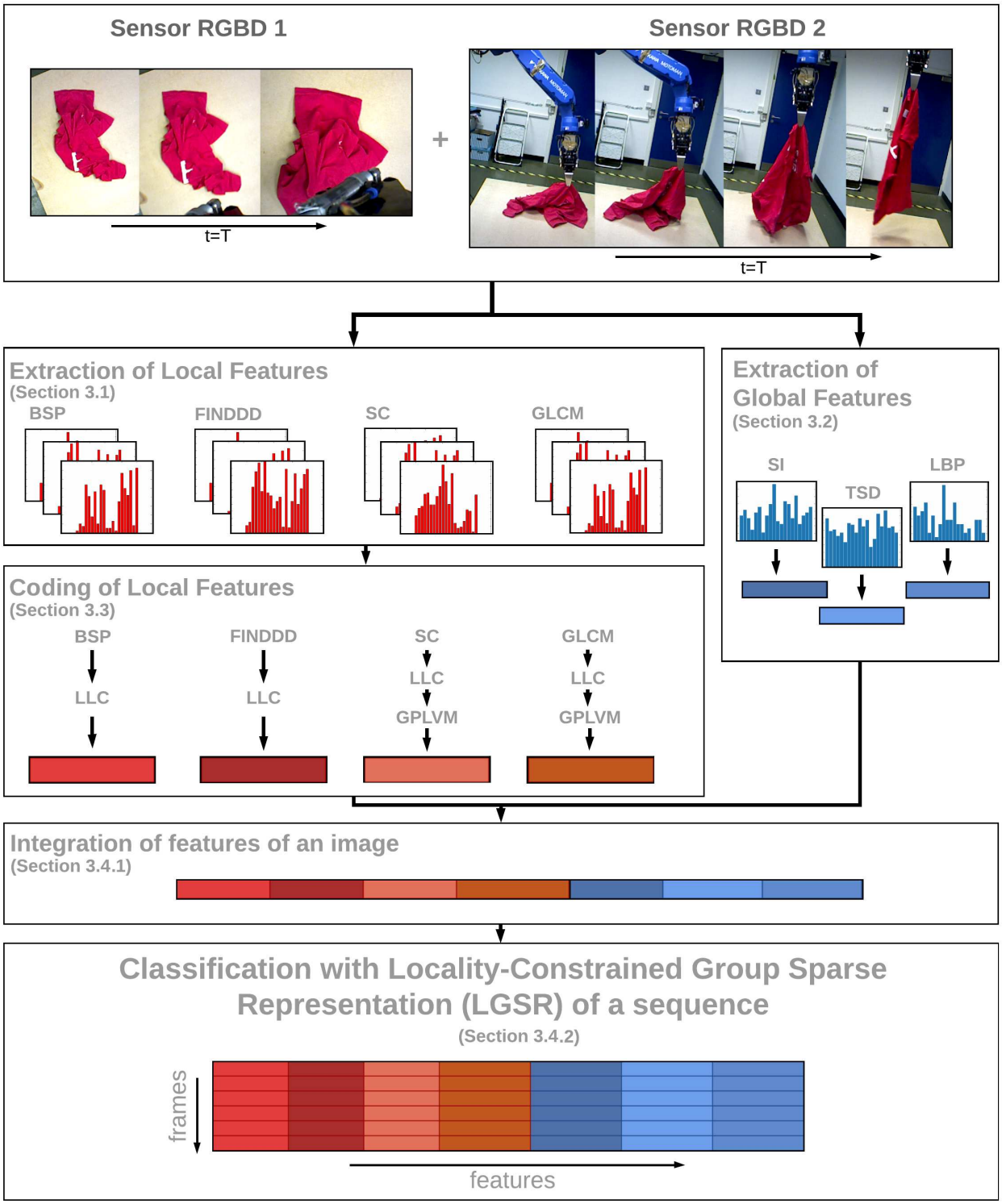

Figure 2: Continuous perception: Local and global features are concatenated into a single feature vector before passing it to the LGSR classification step.

3.2. Visual Features

Local features contribute to describing wrinkles information, a characteristic that possesses only deformable objects and it is a distinctive feature to describe fabric materials as pointed out by [6]. Global features contribute to shape information and are employed to describe shapes while observing how deformable objects change over time. In this work, we found that wrinkles are indeed key to capture the dynamic interaction of the object's particles and to create relationships between particle's locations frame-by-frame, see Section 4.2 – similar observations have already been made in [6] for single-frame recognition and classification. We carry out a comprehensive study on the contributions of each local and global features in Section 4.3, demonstrating its contribution to the continuous visual classification task.

We have thus used the B-Spline Patch (BSP), Histogram of Topology Spatial Distances (TSD), Shape Index (SI) and Histogram of Local Binary Patterns (LBP) as the visual features for local and global feature representations. We refer the reader to [6, 3] for more details about these techniques. Similarly, we describe briefly below the Fast Integral Normal 3D descriptor (FINDDD), Shape Context (SC) and Grey Level Co-occurrence Matrix (GLCM) feature extraction techniques for completeness in our experimental setup.

FINDDD [19] represents the distribution of orientations of the 3D normals in a region around a point of interest in a structured point cloud. The computation of the FINDDD descriptor is based on computing the normal vector for every point in the cloud, using integral images to accelerate the process. Then, the point cloud is divided into sub-regions, and for each sub-region, a descriptor is computed by constructing normal orientation histograms. Instead of using bins defined as angles in spherical coordinates, FINDDD features are distributed regularly across the entire semi-sphere in Cartesian coordinates. The latter avoids concentration around the north pole (maximum elevation), and the uneven area assigned to each bin caused by the angular representation. In this paper, we use the Point Cloud Library [32] implementation to estimate the normals of every point of a structured point-cloud as the basis to compute FINDDD descriptors.

SC[1] [33] describes the relationship between one point with respect to the other points on the shape. This descriptor determines the relationship using a logarithmic-polar distance and classifies these values into a histogram of 12×5 bins. The Shape Context descriptor gives a discriminative global characterization of the shape into a local descriptor since the distances are calculated with respect to other points in the shape. SC, therefore, describes structures in terms of a translation invariant descriptor.

The GLCM [34, 35] technique determines the pixel relationship with other pixels in terms of distance and angle. GLCM calculates the co-occurrence matrix by calculating how often a pixel with a gray-level (grayscale intensity) value occurs in any of the eight defined directions (0, 45, 90, and 135 degrees). Although their analysis methods, the GLCM algorithm is one of the commonly adopted techniques for finding texture information in images of natural scenes and performs well in object recognition [36]. We used Matlab's functions[2] in our experiments. The SVD (singular value decomposition) is calculated from the co-occurrence matrix generating three matrices (U, S and V). U and V represent the left and right singular vectors of the image matrix, and S is a diagonal matrix with singular values. Then L^1 normalization is applied in the diagonal matrix, using this value as the descriptor value.

3.2.1. Distinctive Features

In our experimental design, we also evaluated the integration of distinctive features, such as the collar of jeans, the eyes of a teddy bear, to name a few. Our implementation is based in the Viewpoint Feature Histogram (VFH) descriptors [37] in a selected region performing matching with the k-nearest neighborhood as demonstrated in [38] for grasp point detection in clothing. The VFH descriptor represents four different angular distributions of surface normals in a compound histogram. We use PCL's implementation, where each of these four histograms has 45 bins, and the viewpoint-dependent component has 128 bins, totaling 308 bins. To determine distinctive features, we marked the region where distinctive features appeared in our database and train a naive K-Nearest Neighbour with VFH descriptors. We then search these features over the input image for classification and detection.

For the training phase, we computed the local maximums of an entropy filter over the input depth image to extract potential contours on the deformable object. We used active contour models [39] to select a contour

[1] We use the author's Shape Context implementation which can be found here: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/code/sc_demo/

[2] https://www.mathworks.com/help/images/ref/graycomatrix.html

and describe the part to be detected. The active contours method consists of curves defined within an image domain that can move under the influence of internal forces coming from within the curve itself and, also external forces computed from the image data. The internal and external forces are defined so that the snake will conform to an object boundary or other desired features within an image. Hence, we annotate a distinctive feature as the active curve that describes a specific part of the deformable objects in order to compute VFH descriptors on the selected contour in the depth image. For the classification phase, we follow the same methodology, but we find the ten closest VFH descriptors with respect to the input VFH descriptors. In the case where two or more classes have the same voting, the distances of the neighbors belonging to those classes are added, and we select the shortest distance.

3.3. Visual Feature Coding Techniques

We use the Locality-constrained Linear Coding (LLC) [40, 6] because it has shown to perform more effectively in object and clothing recognition benchmarks. In this paper, we apply this coding technique for each of the local features (BSP, FINDDD, SC, and GLCM) as it can be seen in Figure 2.

We also adopted the Gaussian Process Latent Variable Model (GPLVM) [41] to compress information provided by the local features. That is, GPLVM is non-linear dimensionality reduction technique that generalizes principal component analysis, and it provides a nonlinear mapping to reproduce transformed samples from a latent variable space to an observation space by imposing a Gaussian process prior over the mapping function. This coding technique is used in the local features: SC and GLCM, after the LLC coding technique. The latter can be seen in Figures 2.

3.4. Classification

3.4.1. Feature integration

In this paper, feature integration combines multiple observations of a sequence for recognition and classification. First, for each depth image, we generate a Composite Feature Vector (CFV, see Figure 3) by concatenating each visual feature extracted from the depth image. Then, all feature vectors in the sequence are integrated to create a representation matrix of $n \times F$, where n is the number of views in a sequence and F, the size of the composite features vector v (see Figure 2).

3.4.2. Locality constrained group sparse representation

The Locality Constrained Group Sparse Representation (LGSR) [31] is a classification method commonly used for human gait recognition, where it is needed to classify each input sequence with the information of multiple frames. This method imposes the weighted mixed-norm penalty on the reconstruction coefficients in order to enforce both group sparsity and local smooth sparsity constraints. Thus, LGSR utilizes the intrinsic group information effectively from multiple images within each sequence, treating each test/training sequence as a group of features that combines specific features in an image for classification. In this paper, LGSR provides us with the ability to combine and fuse visual information about the deformable object as it changes over time.

Let $V = [V^1, V^2, ..., V^M]$ and $V^i = [v_1^c, v_2^c, ..., v_n^c]$ where V^c is the cth sequence in the training set and $v_i^c = [L_{BSP}, L_{FINDDD}, Z_{GLCM}, Z_{SC}, S, LBP, TSD]$ is the composite features of the ith view in the cth sequence; n and M are the total numbers of views and sequences in the training set, respectively. We also define the test sequence $Y = [y_1, y_2, ..., y_n]$, where $y_i = [L_{BSP}, L_{FINDDD}, Z_{SC}, Z_{GLCM}, SI, LBP, TSD]$ is the composite features of the ith view in the input sequence. Let us now represent the reconstruction coefficient as $S = [(S^1)^T, (S^2)^T, ..., (S^M)^T]$, where S^c is the reconstruction coefficient for the input sequence with respect to the cth sequence. LGSR thus allows us to enforce group sparsity and local smooth sparsity constraints by minimizing the weighted $l_{1,2}$ mixed-norm-regularized reconstruction error as follows:

$$\begin{aligned} S^* &= argmin_S G(S) \\ &= argmin_S \tfrac{1}{2}\|Y - VS\|_F^2 + \lambda \sum_{c=1}^{M} \|D^c \odot S^c\|_F \end{aligned} \quad (1)$$

where $R(S) = \frac{1}{2}\|Y - VS\|_F^2$ represents the reconstruction error of the input sequence Y with respect to all the training set V. The second term is the weighted $l_{1,2}$ mixed-norm-based regularizer of the reconstruction coefficient S, and $\lambda > 0$ is the regularization parameter to balance these two terms.

$D^c \in \mathbb{R}^{n_c \times n_p}$ is the distance matrix between the views of the cth gallery sequence and the views in the input sequence. To calculate D^c, we compute the distance d_c between the input sequence and the cth gallery sequence using the single-level Earth mover's distance-based temporal matching method [42], and d_{min} with the minimum distance of $d_c|_{c=1}^{M}$.

For the ith composite features from the cth sequence in the training set, we define $D_{ij}^c = exp[(d_c - d_{min})/\sigma]e_{ij}$,

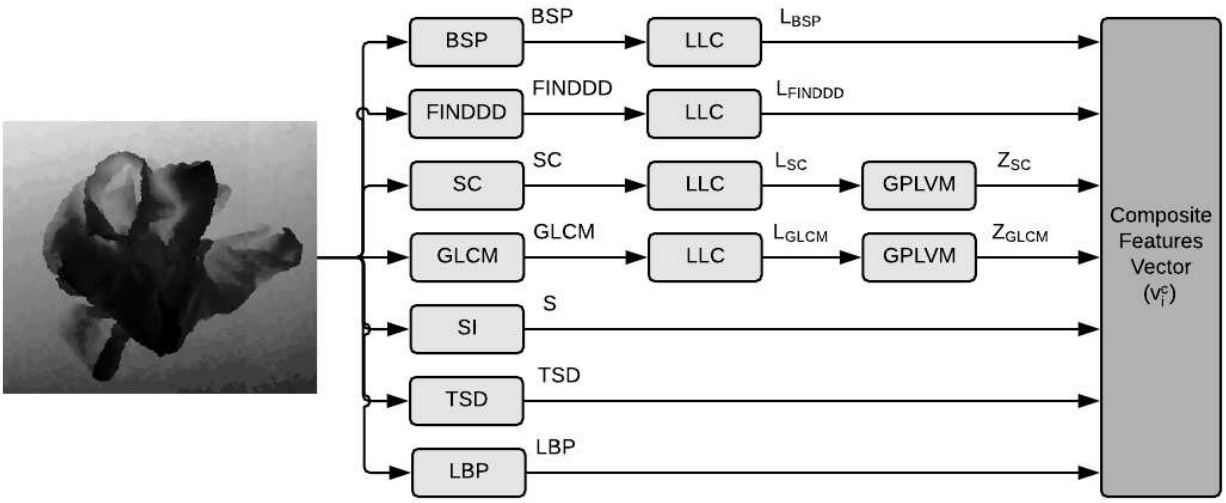

Figure 3: Composition of the final Composite Feature Vector (v_i^c). This vector encodes local and global features to characterize dynamic interactions of deformable objects in a sequence.

where e_{ij} is the Euclidean distance between v_i^c and y_j, and σ is the bandwidth parameter ($\sigma = 1/40, 1/8, 1/4, 1/2$) – we found that the convergence time decreases with a lower value.

$$D_{ij} = exp[(d_c - d_{min})/\sigma]e_{ij} \tag{2}$$

We use the active set-based sub-gradient descent algorithm to solve equation (1), as in [43]. The values of S^c are updated at iteration $t+1$ by equation (3). As follows:

$$S^c_{t+1} = S^c_t - \beta_t \left.\frac{\partial G(S)}{\partial S^c}\right|_{S=S_t} \tag{3}$$

where $\partial G(S)/\partial S^c$ is the updating direction and β_t is the step size determined by a standard line search method. By taking the sub-gradient of $G(S^c)$ with respect to S^c, the updating direction is defined as:

$$\frac{\partial G(S)}{\partial S^c} = \frac{\partial R(S)}{\partial S^c} + \lambda \frac{\partial \|D^c \odot S^c\|_F}{\partial S^c} \tag{4}$$

where,

$$\frac{\partial R(S)}{\partial S^c} = (V^c)^T(VS - Y) \tag{5}$$

$$\frac{\partial \|D^c \odot S^c\|_F}{\partial S^c} = \begin{cases} \frac{D^c \odot D^c \odot S^c}{\|D^c \odot S^c\|} & if\ S^c \neq 0 \\ Z^c & otherwise \end{cases} \tag{6}$$

The particular optimization algorithm is summarized in Algorithm 16. We initialize S as a matrix with all its elements as zero such that all the sequences are added into the active set in order to update the corresponding reconstruction coefficients. After obtaining the optimal reconstruction coefficient S^*, we use the Minimum Reconstruction Error (minRE) criterion to classify the input sequence. We compute the reconstruction error for each class as follows:

$$R_c((S^c)^*) = \frac{1}{2}\|Y - V^c(S^c)^*\|_F \tag{7}$$

where the reconstruction coefficient $(S^c)^*$ is from S^* that corresponds to the cth gallery sequence. Then, we classify the input sequence to $c^* = argmin_c R_c((S^c)^*)$.

4. Experiments

Our working hypothesis is that continuous perception provides robots with greater visual competences for deformable objects. To demonstrate this, we devised a continuous perception approach that used the information obtained from observing the dynamic interactions of different fabrics of garments while a robot picks them up. Therefore, our experimental design consists of performing clothes classification with two different databases.

With these databases, we can evaluate and compare the performance of techniques for deformable objects classification with other approaches from the literature. Figure **??** depicts our proposed continuous perception

Algorithm 1: Optimization Algorithm of LGSR

Input : Y: input sequence, V: training set

Initialize $t = 1, S_t = 0 \in \mathbb{R}^{n \times n_p}, A = \{\}$

Compute D^c between the cth sequence in the training set and the input sequence, $\forall c \in \{1, ..., M\}$.

while $t < T_{Max}$ **do**

  Compute $L_c = \|\partial R(S)/\partial S^c\|_F|_{S=S_t} \forall c \in \{c|S_t^c = 0\}$

  Find $c^* = argmax_c L_c$. **If** $L_{c^*} > \lambda min(D^c)$ **then** $A = c^* \cup A$

  for *each* c *in* A **do**

    Update S_{t+1}^c by using eq. (3) with line search.

    if $S_{t+1}^c = 0$ **then**

      remove c from A

    end

  end

  if $\|S_{t+1} - S_t\|_F < \epsilon (\epsilon = 0.001)$ **then**

    exit WHILE

  end

  t = t+1

end

Output: S

pipeline within a robotics sorting task. We also perform an ablation study to examine the effectiveness and contributions of different visual features (Sections 3.2), and coding (Section 3.3) and classification (Section 3.4) techniques used in our approach.

4.1. Materials: Clothing Databases

For clothing classification, we have collected a large database of RGB-D video sequences of clothing items using two Asus Xtion Pro Live sensors located in the wrists of a dual-arm industrial robot. Then, an existing database was used for making single-shot classifications with different resolutions, to compare the performance between using a high-resolution stereo device and an Asus Xtion Pro Live device (Section 4.5).

First, for continuous perception experiments, we have collected a database of RGB-D video sequences [7][3]. This database features a collection of 'rosbags'[4] containing color and depth images, point clouds, camera information, and all the robot kinematic transformations during the video sequence. Specifically, the database consists of 15 clothing items of 5 categories: t-shirts, shirts, sweaters, jeans, and towels. Each item of clothing is captured from 10 different random configurations, totalling 150 garment videos in random configurations and as being manipulated by our robot. Each sensing device saved RGB-D video streams at 30 Hz. This dataset allows comparisons to be made from different visual views, e.g., at the table, hanging or continuous movement from the side of the robotic action and top-down view.

Second, for single shot experiments, we use the free-configuration clothing database [5]. This database comprises 50 clothing items of 5 categories: t-shirts, shirts, sweaters, jeans, and towels of clothing are captured in 21 different random configurations high-resolution stereo robot head system and an Asus Xtion Pro Live. In total, the database has 1,050 garment images in random configurations for each sensing device; providing for each clothing item an RGB image, depth image, and segmented mask. Furthermore, each of the images has a 16 MegaPixels (4928×3264) image resolution from a stereo robot head [6] and a VGA (640×480) image resolution from an Asus Xtion Pro Live.

4.2. Continuous Perception Experiments

We evaluated our approach on the RGB-D video sequences from our continuous clothing database with two methods of the state-of-the-art: interactive perception [3] and single-shot perception [6]. Since these two methods only evaluate images, three representative images were selected: the first image (when the object is on the table), the last image (when the object is hanging from a gripper) and the image with the best result. For the third case, the methods evaluated each image of the sequence and the result with the best performance was selected. Table 4 shows the comparison between our approach and these two methods using the three representative images.

Classification accuracy results can be depicted in Table 4 and Figure 4. Overall, the proposed approach observes a mean classification accuracy of 66.7%, with specific-class accuracy of 58.0%, 41.6%, 83.8%, 67.0% and 83.8% for the t-shirt, shirt, sweater, jeans and towel classes, respectively. From the results, we noticed that the sweater and towel classes represent the best classification scores due to the inter-class dissimilarities in shape and surface typologies. Although the sweater class gets the best classification scores, this class has higher false positives, resulting in a lower score for the shirt class. This reduction in performance is because deformations in a sweater and shirt classes are similar since both classes have the same fabric material, i.e.,

[3] Available at http://dx.doi.org/10.5525/gla.researchdata.669

[4] http://wiki.ros.org/rosbag

Table 1: Performance comparison between our proposed method with two methods of the state-of-the-art. First, the interactive perception method[3], with the features LBP, SI and TSD (L-S-T) using Gaussian Processes (GP). Second, the single-shot perception[6], with the features LBP, SI, TSD and BSP (L-S-T-B) using support vector machine (SVM).

Algorithm	accuracy
L-S-T with GP (first image)	35.6%
L-S-T with GP (last image)	37.47%
L-S-T with GP (best image)	35.00%
L-S-T-B with SVM (first image)	38.93%
L-S-T-B with SVM (last image)	37.67%
L-S-T-B with SVM (best image)	39.40%
Our Method	66.7%

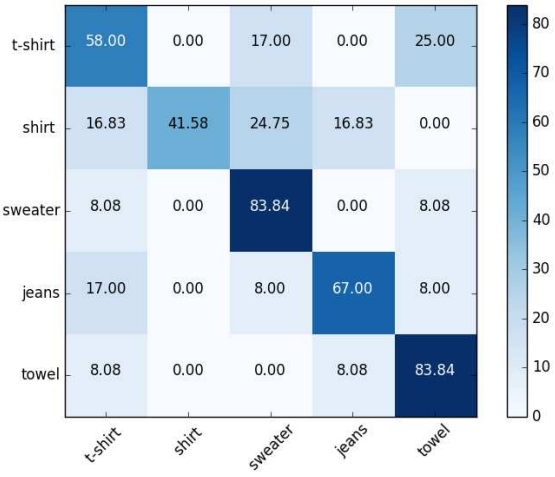

Figure 4: Confusion matrix of our method using our continuous clothing database.

cotton based fabric. For the interactive perception approach using our continuous database, the best average accuracy score is 37.47%, with 41.7%, 17.3%, 31.7%, 32.0%, 51.3% for the classes t-shirt, shirt, sweater, jeans, and towel respectively. Similarly, the best average accuracy with the single-shot perception approach is 39.4%, with 48.0%, 16.3%, 25.0%, 55.0% and 52.7% for the classes t-shirt, shirt, sweater, jeans, and towel, respectively.

By considering the average accuracy and individual accuracy of each class, we can confirm that our continuous perception approach outperforms the interactive perception approach by 31.9%, and 29.23% for the single-shot perception. We can thus conclude that our approach improves the capabilities of a robotic garment sorting task, specifically those tasks that consist of manipulating highly deformable objects since the object space is no longer described based on the 3D structure of its visible surface but by observing how the garment's fabric changes over time.

4.3. Ablation study.

To investigate how different visual features approaches contribute and perform in our continuous perception approach, we carried out ablation experiments as listed in Tables 2 and 3. The experiments in Table 2 are about evaluating the effectiveness of local and global features for the continuous perception clothing classification task over different configurations. That is, we deactivated different features and divided these experiments as follows: proposed method (ID 1.1), local features (IDs 1.2-1.5), global features (IDs 1.6-1.8), only global features (ID 1.9) and only local features (ID 1.10). In ID 1.2 - 1.8, we deactivated the contribution of one visual feature, while leaving the rest unchanged. Similarly, Table 3 shows the experiments that evaluate the impact of the coding algorithms LLC and GPLVM for the classification task. These coding algorithms are applied only on local features. The experiments are distributed in the following way: the proposed method (ID 2.1), only coding using LLC (IDs 2.2-2.5) and only coding using GPLVM (IDs 2.6-2.7).

Figure 5 depicts the results of the experiments described in Table 2. As observed in Figure 5, local features (ID 1.2 - 1.5) capture more distinctively the dynamic interactions of clothing particles. This is because classification scores are close to or below 50% classification score, lower than when one global feature is not considered. The latter is further supported by the classification scores obtained in Figure 6. Notably, the contributions of BSP and SC local features have a considerable impact in the classification scores, since when either of them are not considered, classification scores are below 30% but, when fused without FINDDD, the classification score is close to 60% (see ID 2.3 in Figure 6).

We also discovered that the GPLVM coding technique (ID 2.6 and 2.7) does not contribute considerably to the continuous classification task with respect to LLC, so it is considered an optional technique to decrease the computational load. We, therefore, deduce that LLC captures the most distinctive features. Global features observe minimal contributions, e.g., classification scores for ID 1.9 and 1.10. Even though the combination of local and global features represents the best classification score, global features only contribute to approximately 3% of the total score. That is, local visual features characterize unique information about the

Table 2: Experiments for ablation studies of the features of the proposed solution. Where 'yes' indicates that the feature is activated and 'no' when it is disabled.

ID	BSP	FINDDD	SC	GLCM	SI	LBP	TSD
1.1	yes	yes	yes	yes	yes	yes	yes
1.2	no	yes	yes	yes	yes	yes	yes
1.3	yes	no	yes	yes	yes	yes	yes
1.4	yes	yes	no	yes	yes	yes	yes
1.5	yes	yes	yes	no	yes	yes	yes
1.6	yes	yes	yes	yes	no	yes	yes
1.7	yes	yes	yes	yes	yes	no	yes
1.8	yes	yes	yes	yes	yes	yes	no
1.9	no	no	no	no	yes	yes	yes
1.10	yes	yes	yes	yes	no	no	no

Table 3: Experiments for ablation studies of the coding algorithms of the proposed approach. Where 'yes' indicates that the feature is activated and 'no' when it is disabled.

ID	LLC				GPLVM	
	BSP	FINDDD	SC	GLCM	SC	GLCM
2.1	yes	yes	yes	yes	yes	yes
2.2	no	yes	yes	yes	yes	yes
2.3	yes	no	yes	yes	yes	yes
2.4	yes	yes	no	yes	yes	yes
2.5	yes	yes	yes	no	yes	yes
2.6	yes	yes	yes	yes	no	yes
2.7	yes	yes	yes	yes	yes	no
2.8	yes	yes	yes	yes	no	no

dynamics of the fabrics while global feature captures the overall shape of clothing as it is being pulled.

4.4. Continuous Perception Strategy

To evaluate how many images should be considered for each sensor, we determine the number of images needed to be passed to our approach to achieving the classification scores described in previous sections. These results can be observed in Table 4. Sensor 1 corresponds to the RGBD camera on the arm that manipulates the garment and captures the first images of the garment on the table, i.e., an egocentric view. Sensor 2 is the RGBD camera that captures images from a distance while the other arm picks up the garment from the table, i.e., exocentric view.

Table 4 shows that it is better to use 1 or 2 images of sensor 1 that has a view from above of the garment on

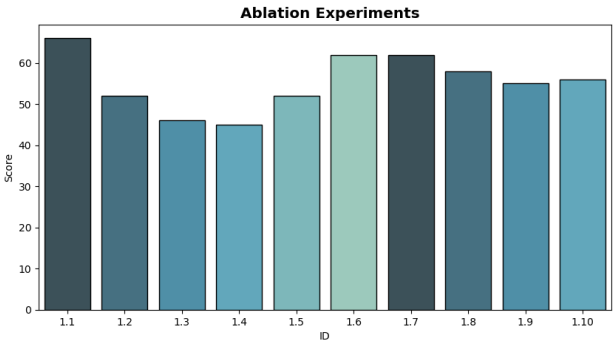

Figure 5: Ablation study results of the experiments shown in the Table 2.

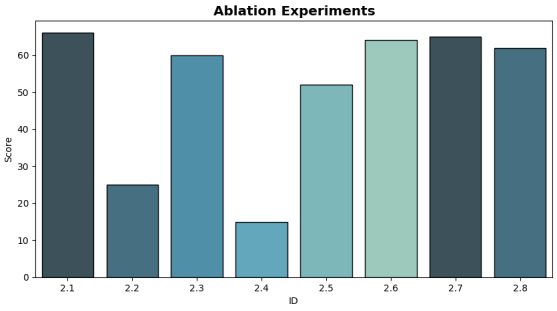

Figure 6: Ablation study results of the experiments shown in the Table 3.

Table 4: Sensor RGBD 2 and Sensor RGBD 1

Sen. 2 \ Sen. 1	1	2	3	4
5	55.0%	55.0%	55.0%	50.0%
15	66.7%	66.7%	63.3%	53.3%
25	58.3%	61.7%	61.7%	51.7%

the table. Also, we found that our approach achieves a better performance while using the last fifteen images of the sensor 2 – this corresponds to when the garment is almost hanging from the robot's gripper, and it is not crumpled.

4.5. Single-Shot Experiments

In order to compare the performance of the features and the coding techniques used in our proposed approach with the state-of-the-art approaches, we validated our approach using the free-configuration clothing database. This database is about evaluating single-shot classification and does not feature video sequences to explore our continuous approach fully. For this, we replace the LGSR classification method (Section 3.4.2) with an SVM classifier but using the same local features (BSP, FINDDD, SC, and GLCM) and global features (SI, LBP, and TSD). In order to increase the accuracy of our approach, we also evaluated the integration of distinctive features, specifically the collar and waist information in the shirt and jeans classes respectively as in Martinez et al. [38]. The collar and waist information is based in VFH descriptors [37], a compound histogram with four angular distributions of surface normals in the selected region.

For these experiments, we only used the images with high resolution (4928×3264), and we obtain an average classification score of 84.8% (see Table 5) using our approach without LGSR. As shown in the confusion matrix of the Figure 7(a), the classification score for each class is: 91%, 67.1%, 83.8%, 90.9% and 91.1% for the t-shirt, shirt, sweater, jeans and towel classes, respectively. These values can be compared with the best result of the state-of-the-art [6] in this database, with an average classification score of 83.2% (see Table 5) and with individual scores of: 89.2%, 70.0%, 80.8%, 87.0% and 88.8% (c.f. Figure 6(f) in [6]) for the t-shirt, shirt, sweater, jeans and towel classes, respectively. We must note that the improvement is only marginal but allow us to confirm that our vision approach is comparable with current approaches to clothing classification while performing single-shot recognition.

For the experiments where we included the collar and waist features, our approach observed an improved average classification score of 87.7% (see Table 5), Figure 7(b) and that class-specific classification scores are: 88.3%, 79%, 87.4%, 93.1% and 90.5% for t-shirt, shirt, sweater, jeans and towel, respectively. By integrating more distinctive visual features descriptions within our approach, we can observe an increase in performance in the classification scores (approx. 3%). We can thus speculate that visual features such as buttons, collars, waists, and so forth, on garments would lead to less inter-class similarities and, consequently, increase class-specific classification scores.

Table 5 also shows a comparison between image resolutions. This allowed us to evaluate if our approach observes a decrease in performance while using different sensing capabilities. These experiments are motivated by the fact that the above classification scores improve while using high-resolution images. Hence, we can observe in Table 5 an increase of 1.1% and 9.4% in low resolution and an increase of 1.6% and 4.5% while using high-resolution images. The latter demonstrates that our approach outperforms results from the state-of-the-art [6] in this database. The increase in performance while using the distinctive features is because the collar of the shirt and waist of the jeans improve the accuracy in these classes and decrease the false positives with respect to other classes.

We must note that these distinctive features resulted in an increase in performance for single-shot recognition/classification tasks. The reason for this is that while perceiving continuously clothing items as our robot picks them up, distinctive clothing features disappear and appear randomly between frames; thus making LGSR to lose accuracy since it usually stayed in a local minimum. Thus, we did not include these distinctive features in the continuous perception experiments in Sections 4.2 and 4.3.

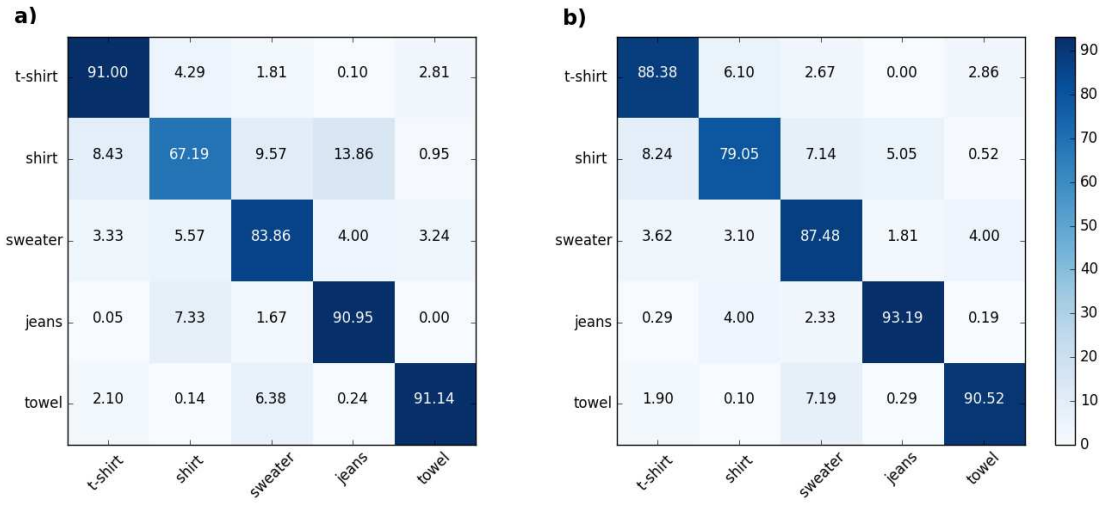

Figure 7: Confusion matrix of the proposed method in the high-resolution Clothing Dataset. **a)** results of the proposed method and **b)** results of the proposed method with distinctive features.

Table 5: Performance comparison between our approach with and without distinctive features against two methods of the state-of-the-art. First, the interactive perception method [3], with the features LBP, SI and TSD (L-S-T) using Gaussian process (GP). Second, the single-shot perception [6], with the features LBP, SI, TSD and BSP (L-S-T-B) using support vector machine (SVM).

Algorithm	Low resolution	High resolution
L-S-T with GP[3]	58.5%	70.8%
L-S-T-B with SVM[6]	64.2%	83.2%
Our Method	65.3%	84.8%
Our Method + Dist. Feat.	73.6%	87.7%

5. Conclusions

In this paper, we have presented a continuous perception approach to classifying clothing categories from video sequences. For this, we have used image sequences from multiple RGB-D sensors from highly wrinkled garment configurations. By adopting the LGSR method, a standard algorithm in human gait action recognition, we have demonstrated that continuous perception can potentially allow a robot to dynamically survey the action and provide us with information to successfully classify clothing categories as the robot carries out a garment sorting task. The latter represents a step forward in traditional sense-plan-act approaches that lead to improvements in automating laundry tasks.

For our approach, we have compiled the first fully-annotated database of RGB-D video scans of clothing items. Video sequences start with the clothing item laying on a flat surface and finalize until the garment is hanging from the gripper of the robot. All videos collected comprise the video streams of two Asus Xtion Pro sensors positioned on the wrists of a dual-arm robot. Likewise, we have also collected the kinematic transformations of the robot while manipulating the garments. This database also can be used for evaluating and validating approaches to clothing recognition in the state of the art while garments are on a flat surface, hanging from a gripper and being picked up by a robot (i.e., continuous perception). To the best of our knowledge, this is the first database of this kind, and we hope it encourages progress in perception methods for highly deformable objects.

Our continuous perception approach has been evaluated using two clothing databases. In all the experiments, we can state that our approach performs well for highly deformed garments. That is, our approach has achieved an average accuracy of 66.7% among 5 categories on our continuous perception database. The latter represents an increase of 39.4% of classification score with respect to current approaches to clothing classification and recognition [6, 6]. We also compared the classification performance of our approach with the free-configuration clothing database. Similarly, our proposed approach advances the state of the art with respect to previous garment databases [5]. Results demonstrated that the rigorous fusion of local and global visual features with appropriate coding techniques (in-

formed by the ablation study in Section 4.3) observed increases in classification scores from 64.2% [6] to 73.6% while using low resolution images, and from 83.2% [6] to 87.7% while using high resolution images.

For future work, we propose to incorporate a complex segmentation algorithm to increase the ability of the robot to analyze garments starting from a pile. Also, considering the improvement of integrating distinctive features, it would be possible to improve the classification performance for a robot sorting task, and garment classification.

Our continous perception dataset will allow us to explore deep learning approaches to deformable object recognition and classification to overcome current engineered approaches.

Acknowledgements

Luz Martínez was funded in this work by CONICYT-PCHA/Doctorado Nacional/2014-21140280.

References

[1] T. B. Jørgensen, S. H. N. Jensen, H. Aanæs, N. W. Hansen, N. Krüger, An adaptive robotic system for doing pick and place operations with deformable objects, Journal of Intelligent & Robotic Systems (2018). URL: https://doi.org/10.1007/s10846-018-0958-6. doi:10.1007/s10846-018-0958-6.

[2] I. Mariolis, S. Malassiotis, Matching folded garments to unfolded templates using robust shape analysis techniques., Berlin: Springer, 2013, pp. 193–200. doi:10.1007/978-3-642-40246-3_24.

[3] K. Sun, S. Rogers, G. Aragon-Camarasa, J. Siebert, Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 2464–2470.

[4] V. Petrík, V. Smutny, P. Krsek, V. Hlaác, Robotic garment folding: Precision improvement and workspace enlargement., in: C. Dixon, K. Tuyls (Eds.), TAROS, volume 9287 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 204–215.

[5] L. Sun, G. Aragon-Camarasa, Clopema clothes, University of Glasgow (2016). URL: http://researchdata.gla.ac.uk/270/. doi:10.5525/gla.researchdata.270.

[6] L. Sun, G. Aragon-Camarasa, S. Rogers, R. Stolkin, J. P. Siebert, Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017, 2017, pp. 6699–6706. doi:10.1109/IROS.2017.8206586.

[7] L. Martinez, G. Aragon Camarasa, J. Siebert, J. Ruiz-del Solar, Continuous perception for clothing understanding in robotic applications, University of Glasgow (2018). URL: https://owncloud.gla.ac.uk/cloud/s/FierZW2obV8leFb. doi:10.5525/gla.researchdata.669.

[8] A. Ramisa, G. Alenyá, F. Moreno-Noguer, C. Torras, Using depth and appearance features for informed robot grasping of highly wrinkled clothes., in: ICRA, IEEE, 2012, pp. 1703–1708.

[9] B. Willimon, I. D. Walker, S. Birchfield, Classification of clothing using midlevel layers, in: ISRN Robotics, 2013.

[10] A. Doumanoglou, T.-K. Kim, X. Zhao, S. Malassiotis, Active random forests: An application to autonomous unfolding of clothes., in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), ECCV (5), volume 8693 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 644–658.

[11] Y. Kita, T. Ueshiba, F. Kanehiro, N. Kita, Recognizing clothing states using 3d data observed from multiple directions, in: 13th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2013, Atlanta, GA, USA, October 15-17, 2013, 2013, pp. 227–233. doi:10.1109/HUMANOIDS.2013.7029980.

[12] Y. Li, C.-F. Chen, P. K. Allen, Recognition of deformable object category and pose, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2014.

[13] Y. Li, D. Xu, Y. Yue, Y. Wang, S.-F. Chang, E. Grinspun, P. K. Allen, Regrasping and unfolding of garments using predictive thin shell modeling, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2015.

[14] Y. Kita, N. Kita, A model-driven method of estimating the state of clothes for manipulating it., in: WACV, IEEE Computer Society, 2002, pp. 63–69.

[15] Y. Kita, T. Ueshiba, F. Kanehiro, N. Kita, Recognizing clothing states using 3d data observed from multiple directions, in: 13th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2013, Atlanta, GA, USA, October 15-17, 2013, 2013, pp. 227–233. doi:10.1109/HUMANOIDS.2013.7029980.

[16] B. Willimon, S. Birchfield, I. D. Walker, Model for unfolding laundry using interactive perception., in: IROS, IEEE, 2011, pp. 4871–4876.

[17] B. Willimon, I. D. Walker, S. Birchfield, A new approach to clothing classification using mid-level layers., in: ICRA, IEEE, 2013, pp. 4271–4278.

[18] A. Ramisa, G. Alenyá, F. Moreno-Noguer, C. Torras, Learning rgb-d descriptors of garment parts for informed robot grasping, Engineering Applications of Artificial Intelligence 35 (2014) 246–258. doi:10.1016/j.engappai.2014.06.025.

[19] A. Ramisa, G. Alenya, F. Moreno-Noguer, C. Torras, Finddd: A fast 3d descriptor to characterize textiles for robot manipulation, in: Proceedings of the International Conference on Intelligent Robots and Systems (IROS), 2013, pp. 824–830.

[20] Y. Li, Y. Wang, M. Case, S.-F. Chang, P. K. Allen, Real-time pose estimation of deformable objects using a volumetric approach., in: IROS, IEEE, 2014, pp. 1046–1052.

[21] L. Sun, G. Aragon-Camarasa, W. P. Cockshott, S. Rogers, J. P. Siebert, A heuristic-based approach for flattening wrinkled clothes., in: A. Natraj, S. Cameron, C. Melhuish, M. Witkowski (Eds.), TAROS, volume 8069 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 148–160.

[22] L. Sun, G. Aragon-Camarasa, S. Rogers, J. P. Siebert, Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening., in: ICRA, IEEE, 2015, pp. 185–192.

[23] Y. Li, X. Hu, D. Xu, Y. Yue, E. Grinspun, P. K. Allen, Multi-sensor surface analysis for robotic ironing, CoRR abs/1602.04918 (2016). URL: http://arxiv.org/abs/1602.04918. arXiv:1602.04918.

[24] A. Ramisa, G. Alenyá, F. Moreno-Noguer, C. Torras, Determining where to grasp cloth using depth information., in: CCIA, volume 232 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2011, pp. 199–207.

[25] I. Mariolis, G. Peleka, A. Kargakos, S. Malassiotis, Pose and category recognition of highly deformable objects using deep learning., in: ICAR, IEEE, 2015, pp. 655–662.

[26] A. Gabas, E. Corona, G. Alenyà, C. Torras, Robot-aided

cloth classification using depth information and cnns, in: Articulated Motion and Deformable Objects - 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings, 2016, pp. 16–23. doi:10.1007/978-3-319-41778-3_2.

[27] E. Corona, G. Alenyà, A. Gabas, C. Torras, Active garment recognition and target grasping point detection using deep learning, Pattern Recognition 74 (2018) 629–641. doi:10.1016/j.patcog.2017.09.042.

[28] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59 (2004) 2004.

[29] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. G. Rodríguez, A review on deep learning techniques applied to semantic segmentation, CoRR abs/1704.06857 (2017). URL: http://arxiv.org/abs/1704.06857. arXiv:1704.06857.

[30] B. Leibe, A. Leonardis, B. Schiele, Combined object categorization and segmentation with an implicit shape model, in: In ECCV workshop on statistical learning in computer vision, 2004, pp. 17–32.

[31] D. Xu, Y. Huang, Z. Zeng, X. Xu, Human gait recognition using patch distribution feature and locality-constrained group sparse representation, IEEE Trans. Image Processing 21 (2012) 316–326. URL: https://doi.org/10.1109/TIP.2011.2160956. doi:10.1109/TIP.2011.2160956.

[32] R. B. Rusu, S. Cousins, 3D is here: Point Cloud Library (PCL), in: IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, 2011.

[33] S. Belongie, J. Malik, J. Puzicha, Shape context: A new descriptor for shape matching and object recognition, Advances in Neural Information Processing Systems 13 (2001) 831–837.

[34] R. Haralick, K. Shanmugam, I. Dinstein, Texture features for image classification, IEEE Transactions on Systems, Man, and Cybernetics 3 (1973).

[35] R. W. Conners, M. M. Trivedi, C. A. Harlow, Segmentation of a high-resolution urban scene using texture operators, Computer Vision, Graphics, and Image Processing 25 (1984) 273–310.

[36] M. Saraswat, A. K. Goswami, A. Tiwari, Object recognition using texture based analysis, International Journal of Computer Science and Information Technologies, 4 (2013) 775–782.

[37] R. B. Rusu, G. R. Bradski, R. Thibaux, J. M. Hsu, Fast 3d recognition and pose using the viewpoint feature histogram., in: IROS, IEEE, 2010, pp. 2155–2162.

[38] L. M. Martínez, J. Ruiz-del Solar, Recognition of grasp points for clothes manipulation under unconstrained conditions, In Proceedings of 22th RoboCup International Symposium, Lecture Notes in Computer Science. (2017).

[39] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, International Journal of Computer Vision 1 (1988) 321–331.

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Classification, 2010.

[41] N. Lawrence, Gaussian process latent variable models for visualisation of high dimensional data, in: In NIPS, 2003, p. 2004.

[42] D. Xu, S. F. Chang, Video event recognition using kernel methods with multilevel temporal alignment, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1985–1997. doi:10.1109/TPAMI.2008.129.

[43] H. Lee, A. Battle, R. Raina, A. Y. Ng, Efficient sparse coding algorithms, in: B. Schölkopf, J. C. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems 19, MIT Press, 2007, pp. 801–808.