Visual Grasp Point Localization, Classification and State Recognition in Robotic Manipulation of Cloth: an Overview

P. Jiménez

Institut de Robòtica i Informàtica Industrial (CSIC - UPC)
Llorens i Artigas 4-6, E-08028 Barcelona, Spain
e-mail: pjimenez@iri.upc.edu

Abstract

Cloth manipulation by robots is gaining popularity among researchers because of its relevance, mainly (but not only) in domestic and assistive robotics. The required science and technologies begin to be ripe for the challenges posed by the manipulation of soft materials, and many contributions have appeared in the last years. This survey provides a systematic review of existing techniques for the basic perceptual tasks of grasp point localization, state estimation and classification of cloth items, from the perspective of their manipulation by robots. This choice is grounded on the fact that any manipulative action requires to instruct the robot where to grasp, and most garment handling activities depend on the correct recognition of the type to which the particular cloth item belongs and its state. The high inter- and intraclass variability of garments, the continuous nature of the possible deformations of cloth and the evident difficulties in predicting their localization and extension on the garment piece are challenges that have encouraged the researchers to provide a plethora of methods to confront such problems, with some promising results. The present review constitutes for the first time an effort in furnishing a structured framework of these works, with the aim of helping future contributors to gain both insight and perspective on the subject.

Keywords: deformable object manipulation, robotic vision, clothing, cloth state recognition, garment classification, grasp point localization

1. Introduction

Robots intended to be increasingly versatile should include the capability of manipulating deformable objects. In particular, being able to handle cloth items should become a standard requirement for robots for full deployment in domestic environments, in assistive robotics, in service scenarios like hotels and hospitals, and of course also in industrial laundry or garment manufactures. However, robotic manipulation of fabrics has to face the intrinsic difficulty of dealing with a highly flexible material, thus with an enormously varying appearance. This means that a specific cloth item –a garment piece, for example– exhibits a practically infinite range of possible shapes, from a canonical extended flat shape up to a completely crumpled state, with intermediate states of varying wrinkledness, as well as a vastity of folded states, not to speak from partial or total reversal (parts of the cloth turned inside out). To this intra-garment variability one has to add the im-

mense range of different cloth items appearing in a standard household (inter-garment variability).

Figure 1: (a) Inter-garment variability: A simple room scene with a miscellany of cloth items. Garments display a variety that ranges from little baby socks to an adult's coat. But there are more objects made of cloth in a household, like the eiderdown's cover, the pillowcases or the curtains in the image. (b) Intra-garment variability: The same object –a shirt– in quite different configurations, the shape outline exhibits great variability.

It is quite difficult to predict the exact outcome of a specific action on a piece of cloth. Fabric displays an intricate behavior, related to its anisotropic and nonlinear mechanical response. As pointed out in [1], subtle mechanical actions are amplified into large draping or motion variations. Deformations on cloth typically appear as creases and wrinkles, or even as folds or inside-out turnings. Such bending deformations store quite low elastic energy, which means that in general cloth does not tend to spontaneously recover the shape previous to the deformation. The different fabric types exhibit a diverse mechanical behavior as well, and also a distinctive appearance, beyond the specific colors that depend on the used dyes. This typology depends not only on the material the yarns are made of, but also on how they are woven or knitted. Despite all these difficulties, models are certainly used to try to predict cloth behavior in response to specific actions of the robot. More specifically, they aim at showing which should be the appearance (the shape) of a garment of a specific type in a given state. Classical modeling techniques used in other Engineering fields, mainly in Simulation or Computer Graphics, have also been used in the robotic cloth manipulation context, like mass-spring models in hanging garments state recognition [2] or continuum representations in state recognition for cloth-straightening action selection [3]. Besides, roboticists have developed their own models in specific cloth handling applications, like the *parametrized shape model* [4] (cloth recognition with the item spread out on a table and folding state identification), or the *polygonal model* [5] (also used in garment matching and fold tracking). Figure 2 illustrates the main features of these models.

Models like the described ones are aimed at simplifying the domain of possible continuous deformations by discretization of the number of possible states. For instance, the shape of a given garment type when grasped and hanging at one of the model's nodes is known (e.g. from simulation) and can be compared with the real image of the candidate item after a manipulation and regrasping procedure that ensures the grasp at one of such points. Clearly, such manipulation process has necessarily to be guided by sensory input. Although simple sensors like tactile sensing devices integrated in the gripper's fingertips

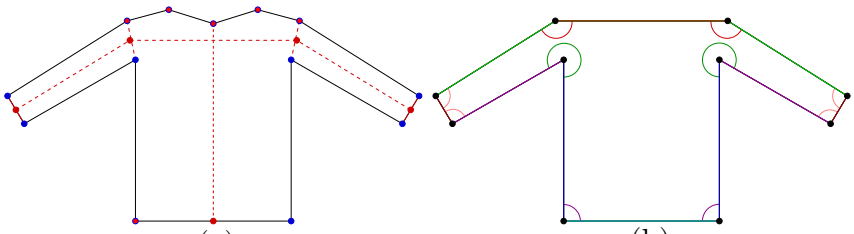


Figure 2: The parameterized shape model [4] (a) and the polygonal model [5] (b) of a long-sleeved shirt. (a) Parameters are the 2D coordinates of the 11 nodes marked in red and the two sleeve widths, also highlighted in red. Landmark points, defining the contour of the garment, are shown in blue, note that some parameter nodes are also landmark points (thus shown in red encircled in blue). Landmark points not in the parameter set are easily computed from the given parameters. (b) Polygonal models are defined by a set of vertices and by their mutual relative positions, in turn determined by the inner angles and the lengths of the segments joining them. These lengths are not absolute distances but relative to the overall perimeter length. Angles and segments measuring the same are drawn in equal colors.

or more sophisticated ones like wrist-mounted F/T sensors can already provide quite significative information, vision is by far the most relevant and complete perception channel for the required information during manipulation. It should be noted that all these perception channels are not mutually exclusive, on the contrary, they are complementary as for providing together a better understanding of the scene. However, this survey focusses strictly on visual processes.

The basic manipulation cycle includes the well-known three phases of grasping, moving and releasing. In the case of cloth manipulation, specific problems appear in each one of these steps, which have been described thoroughly in [6], and are mainly related to the uncontrolled deformation and possible entangling of the manipulated object. With independence of the aimed manipulation goal of a specific task on a cloth item, at some moment visual guidance for successful completion is required , either for determining at which point the cloth item has to be grasped, or to ascertain the degree of completion of the task by assessing the state the cloth item is in. To these two basic visual tasks, a third one has to be added in scenarios where multiple types of clothing are present, namely classification: finding out to which clothing type a specific item belongs to is crucial both for state assessment and for establishing the next actions to perform. These three visual tasks lie at the basis of cloth manipulation, they provide the necessary feedback for the fulfillment of the manipulation goals. On the other hand, they are very specific to the handling of fabrics, they are shaped by the deformability and variability of cloth items. For this reason, and for the growing interest towards the robotic manipulation of such challenging material, the convenience of the present survey seems to be well-grounded. Figure 3 illustrates the expected outputs from the three processes.

These perceptual actions appear recurrently in typical cloth manipulation tasks, as illustrated in Figures 4 and 5.

This survey intentionally leaves out those visual tasks that are too much oriented towards very specific applications. Examples include the continuous image processing needed in seam tracking while sewing [7, 8, 9, 10], vision-based control strategies for folding [11], or optical flow analysis [12] and topological coordinates matching [13, 14] for

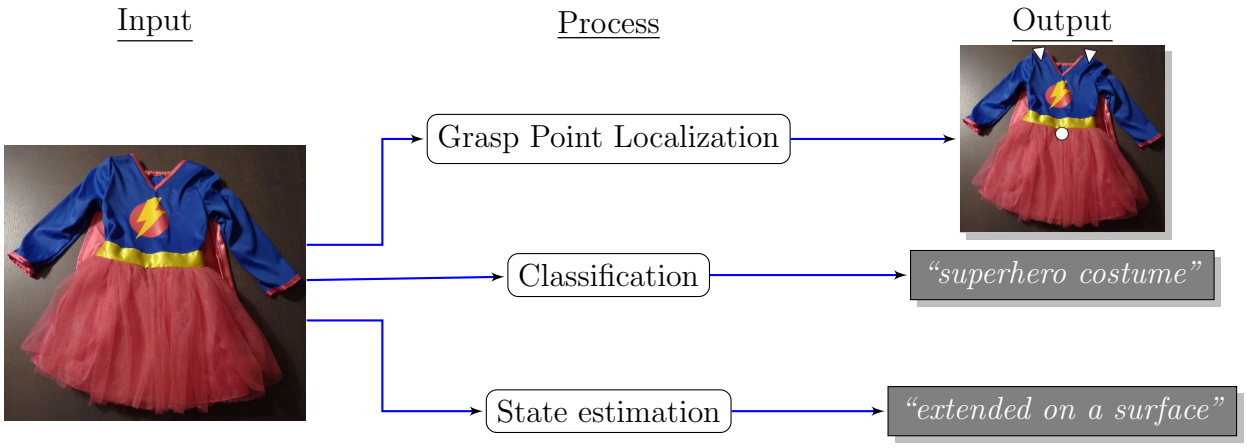


Figure 3: The three visual processes addressed in this survey and their desired outputs. Grasp point localization may provide either a generic point on the cloth item to be grasped (circle), or specific features like the shoulders (triangles). Classification produces a label pointing to a given cloth category, which, depending to the granularity of the categorization, could be as generic as "dress" or as specific as "superhero costume for small girl". State estimation, finally, has been shown here as providing a label, "extended on a surface", that may trigger a subsequent manipulation action like folding.

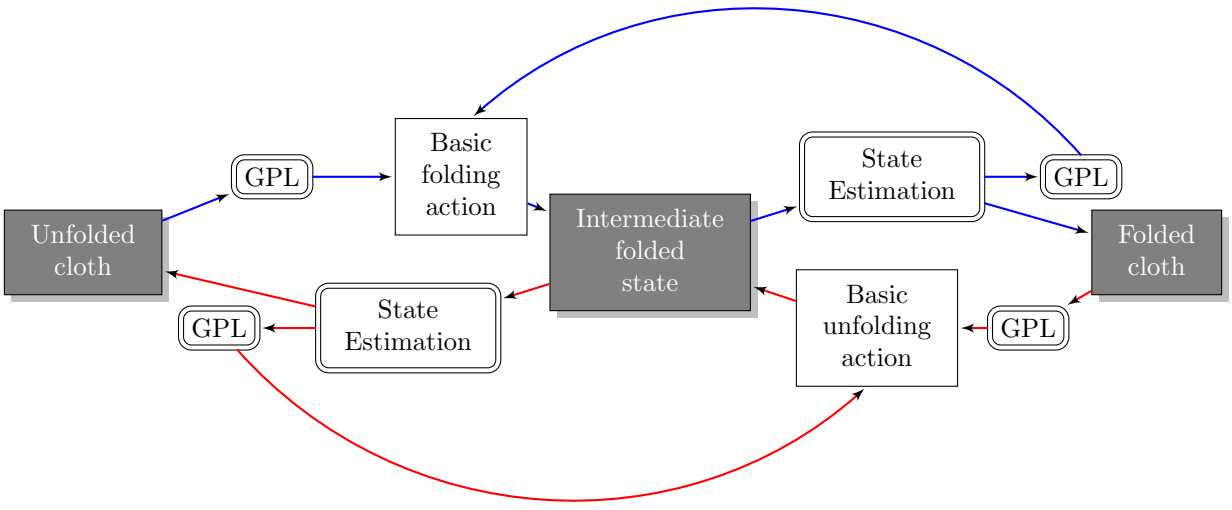


Figure 4: Basic folding and unfolding manipulation tasks, involving the surveyed perceptual actions. The extreme states are assumed to be known beforehand. Furthermore, the type of garment is also assumed to be known (otherwise, a Classification node should be placed before the first folding (or unfolding) action).

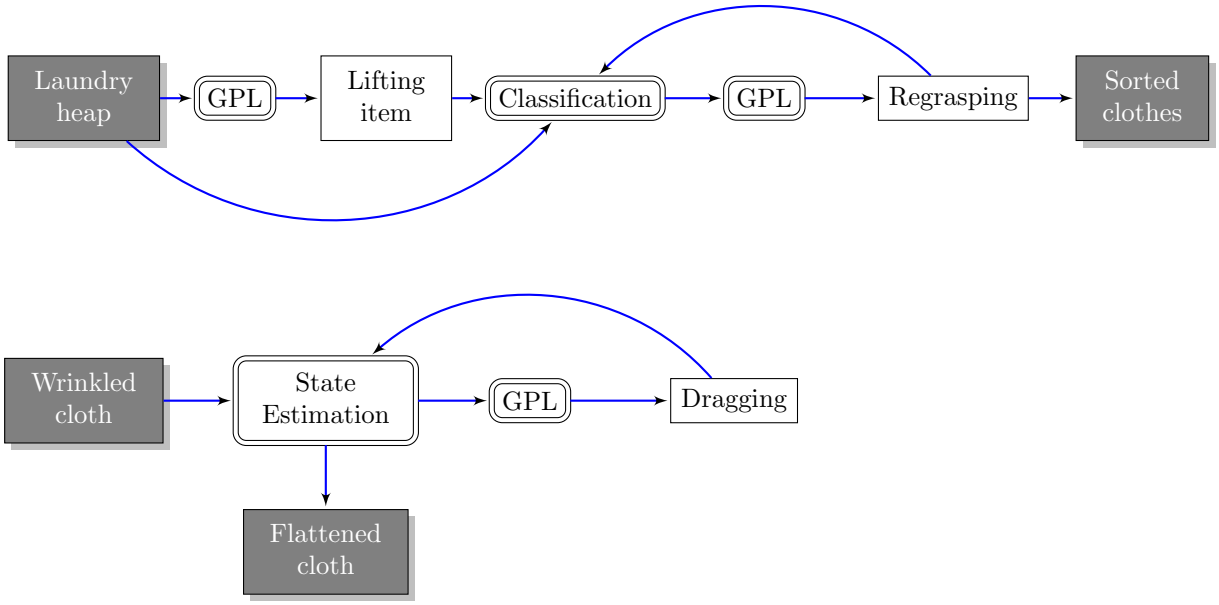


Figure 5: Basic laundry sorting and cloth flattening (by dragging) manipulation tasks, involving the surveyed perceptual actions. In the first, classification may eventually be made without lifting the object, by direct visual inspection. As for the flattening application, the type of cloth is assumed to be already known. Dragging is assumed to include the actual grasping and releasing actions.

the assessment of dressing operations.

Along the survey, it will become clear that the instances of the three addressed visual tasks are quite variable from one work to the other. Nonetheless, the following startpoints and hypotheses may be set:

- Presentation of the cloth part: All the analyzed works specify whether they consider isolated cloth items or appearing among others in a bunch or heap, and whether the cloth is lying on a surface, or grasped and hanging from the robot's gripper.
- Lighting conditions: unless otherwise stated, all the works assume the clothes to be illuminated by ambient light.
- Occlusions: No work considers the cloth item to be partially occluded by any other object than itself (selfocclusions).
- Manipulation conditions: All of the works assume that the cloth is to be fully autonomously handled by robots, and that the image processing takes places without human intervention.

This review of visual methods and algorithms for cloth manipulation is naturally structured in the following sections. First, the elementary cloth-related perceptual tasks like finding a cloth item in a scene or detecting and measuring wrinkles are tackled in Section 2. Then, the first visual goal of identifying adequate points on the cloth to be grasped is analyzed in its two variants of generic (Section 3.1) and specific grasp points (Section 3.2). Cloth type and state recognition are dealt with in Section 4, where

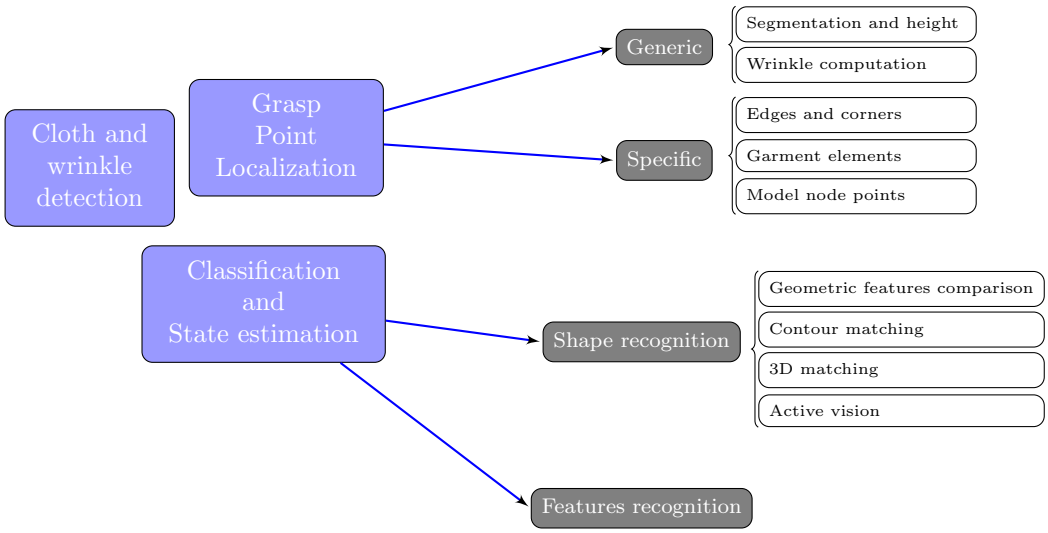


Figure 6: Basic vision tasks for robot cloth manipulation examined in this survey.

recognition methods based on the whole shape (Section 4.1) are distinguished from those based on identifying specific garment features (Section 4.2). Some final remarks are exposed in the Conclusions.

Figure 6 illustrates the layout of the survey.

2. Cloth and wrinkle detection

Most experimental scenarios up to date assume that the cloth part is presented to the vision system. Moreover, in the most common settings, the cloth part is the main object in the image, so that the first operation to perform is just a background segmentation. However, it is reasonable to expect from a robot deployed in a domestic environment to be capable to detect cloth pieces within a more complex scene. This idea inspires the daily assistive robot described in [15], which scans domestic environments in search of clothes. The aim is challenging, as the authors want to identify clothing parts without prior knowledge about the shape or color of specific garment pieces, just by recognizing fabrics as a material. The difficulty is to determine a relevant feature that helps to distinguish cloth from other materials. The authors claim that wrinkles are characteristic features of cloth and they use Gabor filters to extract them from the image. This alone is not enough, but the fact that fabrics tends to display groups of wrinkles allows to discriminate isolated edges that are also detected by this filter. Filtered images are used for training via Support Vector Machine (SVM) techniques (with 20-dimensional feature vectors). A graph cut method completes the extraction of the non-wrinkled part of the image which does also correspond to the same cloth item. Afterwards, the cloth part is picked up and put into a washing machine, by selecting a grasp point, as described in Section 3.1 [16].

A step further is to analyze the wrinkledness state of a piece of cloth (quantifying wrinkles with different orientations). The same authors take this step by extending their Gabor filters-based cloth detection procedure [17], assuming that the robot can take closer views of the garment. They use a set of Gabor filters with different orientations to detect

irregularities on the fabrics in different directions. Wrinkles can be discriminated from border and cloth overlaps or folds by varying the parameters of the wave profile of the filter: the first have a low frequency coefficient, as compared to the high frequency coefficient of the other features. Cloth detection and wrinkledness state measuring methods are summarized in Figure 7.

This type of knowledge is highly valuable in cloth flattening applications: it provides information on where and how long to brush in flattening-by-sweeping, or where to grasp and how long to drag before release in the case of flattening-by-dragging. Besides the Gabor filtering methods [15, 17], other procedures have been proposed as well. Their target is to obtain a geometric description of the wrinkles, including their size, position and orientation. Whereas [18] base their method exclusively on intensity image analysis, by filtering with a threshold function to find areas with abrupt changes in pixel density (in conjunction with edge enhancement procedures), much more recent work uses also depth [19] by using a high-resolution stereo-based sensor to capture a 2.5D range map of the cloth. After background segmentation, a piece-wise B-spline surface approximation is fitted to the resulting depth-map. Then the surface shape is classified into different types at each point, using its *shape index* [20], from which wrinkles are constructed by joining nearby *ridges* (positive extrema of maximal curvature) and detecting the wrinkle's contour at the boundary of convex and concave surfaces of the garment. Wrinkle quantification, the computation of the height and width of the wrinkle, is performed via the wrinkle's triplets (ridge points and adjacent contour points). Finally, wrinkles are scored according to their volume and after using PCA to determine the main direction of each wrinkle, the flattening strategy is applied. The experiments in [19] show the importance of using high-resolution stereo-vision, as compared to a Kinect-like sensor (Xtion, specifically), both in bimanual and single-arm manipulation, requiring much less iterations to flatten the garment. This work has been extended in [21] by verifying the utility and implementation of suitable feature extraction methods within a repeatable simulated environment. Figure 8 shows the described alternative wrinkle measuring methods.

3. Grasp point localization (GPL)

We assume throughout this survey that handling of cloth occurs via mechanical impactive grasps, which means that the fabric is hold by pressure and friction between the gripper's fingers, acting either on the same side –i.e., holding a loop of cloth– (pinch grasp) or on both sides of the cloth (clamp grasp). The idoneity of a point on a cloth part to be grasped, which may guide its localization process, depends in part on the type of grasp used, as well as on the particular state the cloth is in. For flat lying clothes, if they have to be pinch grasped at any interior point, perception has to ensure the contact of the fingers with the fabrics and enough friction to create the loop of cloth when closing the fingers, which hardly can be done with vision alone. More frequent is the case where the lying cloth part displays some wrinkles, and in such case the vision system has to precisely locate an appropriate wrinkle to be grasped (as explained in Section 3.1). The same is applicable for protruding garment features like polo collars for example. Still on clothes lying on a surface, another possibility is to aim at grasping an edge (or a corner) with a clamp grasp. Here the vision system will have to be able to identify such basic features, whereas grasping itself will require either special grippers [22] or special settings

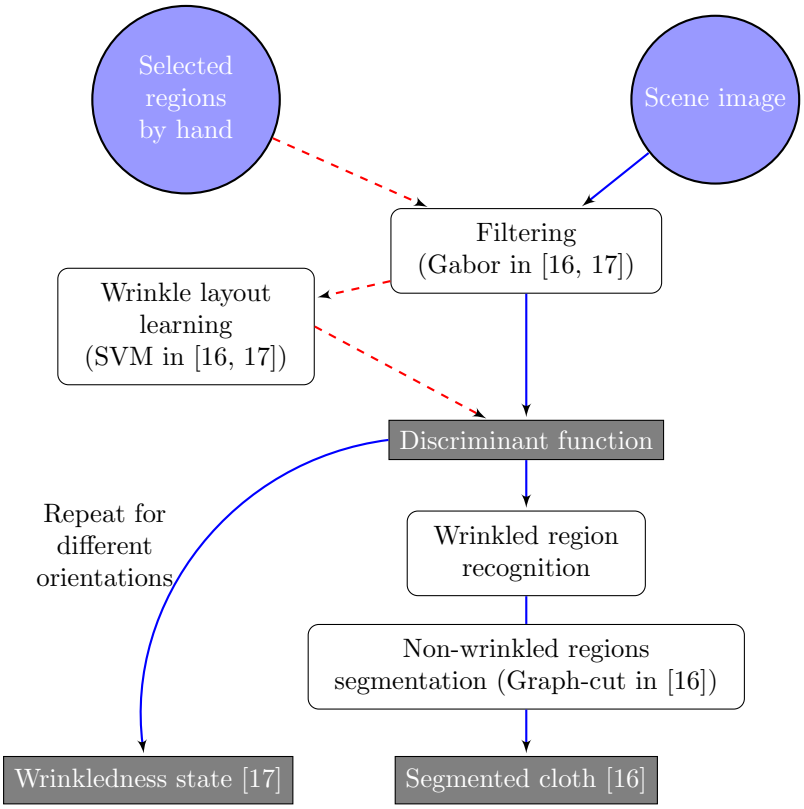


Figure 7: Cloth detection in a scene without prior knowledge. Monocular views are processed in both references.

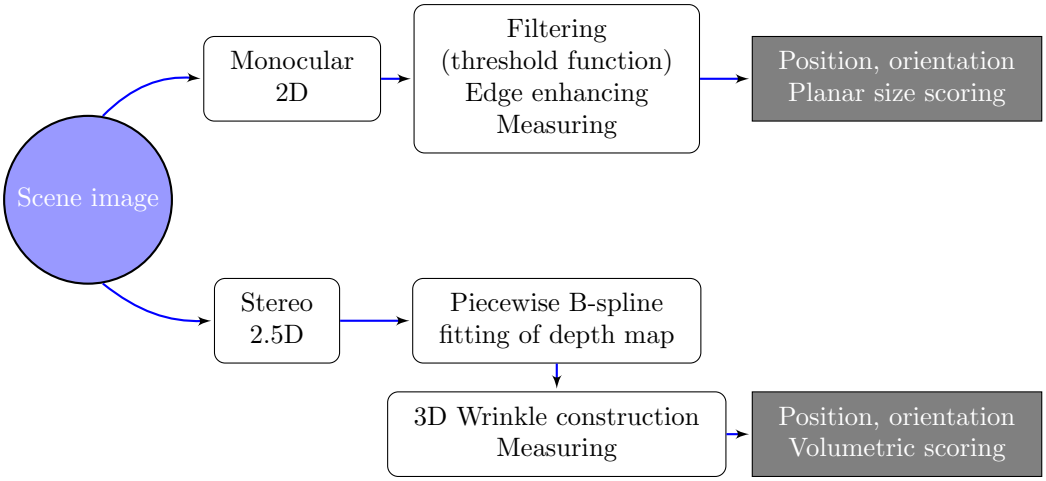


Figure 8: Different approaches for wrinkle measuring. On top, the purely planar procedure of [18], below the volumetric approach of [19].

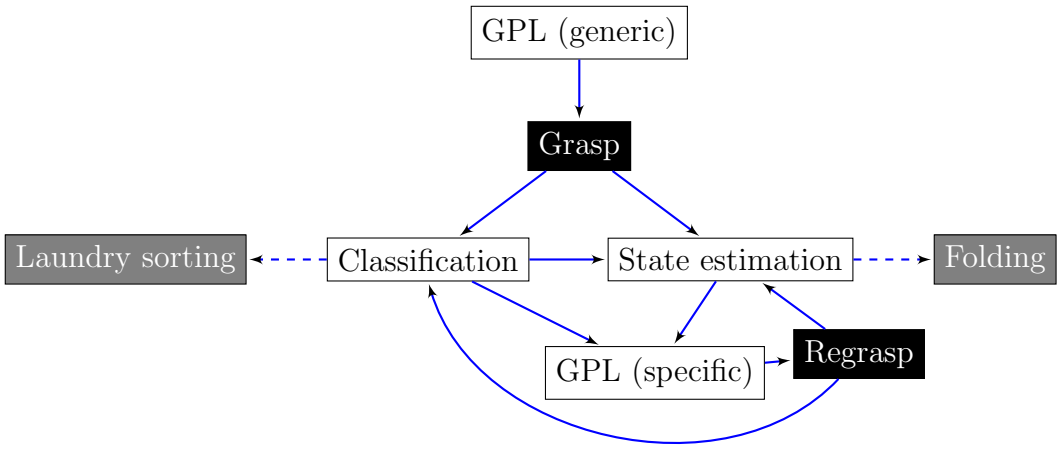


Figure 9: Possible sequences of the reviewed perceptual processes (white nodes), distinguishing the two variants of GPL. The generic GPL may provide a first point for grasping within a state estimation - regrasping (of specific grasp points) cycle. Grasping actions (black nodes) are to be understood as including not only the grabbing of the cloth item at a point, but also some motion of the gripper that presents the grabbed cloth to the vision system. In gray, two possible external manipulation actions are displayed as direct beneficiaries of the classification and state estimation processes. Although displayed in sequence, classification and state estimation can be simultaneous processes, like in [24] (see Section 4.1.1).

like soft foam tables to allow the gripper finger to slide underneath [23]. As for hanging clothes, corners, edges, or specific garment features are aimed for potential grasping, and the vision system has to provide a position and an orientation of the gripper for successful grasping.

Grasp points may be generic or arbitrary, just for picking up the cloth item, or specific, that is, with a precise localization such that the goals of the following manipulation can be fulfilled. In many classification and state estimation procedures, generic GPL constitutes the first step of a sensing - action strategy, where regrasps for further uncertainty reduction are performed on specific grasp points. As in this section the two types of GPL are reviewed, it is suitable to present now the diagram of possible sensing and grasping actions sequences (Figure 9). This diagram has to be interpreted as displaying different alternatives: for example, one sequence could just constitute the path GPL (generic) – Grasp – Classification – Laundry sorting (as a typical garment manipulation application). Two different sequences arise from considering whether state estimation requires a previous classification step (as happens in general) or if this can be avoided by knowing the garment type in advance.

3.1. Generic grasp points

In some applications, the particular point where the cloth piece is grasped is unimportant, as long as a firm grasp on the item is guaranteed. More often, this irrelevance comes from the fact that it is just a *first* point grasp, intended at isolating the cloth part from a bunch of clothes or from the environment, and holding it in such a way that a subsequent regrasp strategy targeting more specific grasp points can be performed. A characteristic

example is picking some garment out from an unordered laundry pile.

This first arbitrary point grasp deserves no special description in some papers like [25]. However, it is far from being a trivial task, and failure to perform it correctly leads either to grasping no cloth at all, or to simultaneously grasp more than one cloth piece, if two or more items are present. The latter may be solved automatically in the case it is a first point grasp within a regrasping strategy, as shown for towel handling in [26], where a subsequent regrasp at a specific corner makes all towels but one fall back on the table. Like in other perceptual processes involving uncertainty in the outcome, action and sensing can be combined in order to cope with the shortcomings of the latter. For instance, in [27] once the grasp point is determined (see below), grasping attempts are made at increasing lower positions over the table, starting just above the estimated height of the item, combined with lateral motions and image processing to check whether the item has been effectively removed.

There are two families of grasp point determination methods, as presented next.

3.1.1. Segmentation and height

Generic point grasp localization is, in first approximation, region-oriented, in contrast to the feature-oriented methods of the next section. More precisely, the aim is to identify a region corresponding to the cloth to pick up, and an arbitrary point within this region is chosen as grasp point. The related visual procedure is clearly image segmentation, and, if only one cloth item is present, this reduces just to **foreground-background segmentation** (or background segmentation for short). Segmentation can be as simple as a binarization to eliminate the background [28, 29], or more sophisticated histogram- or graph-based segmentation procedures aiming at separating different cloth items or to get some information about its surface condition (i.e., its relief). Height –or more generally, depth– information about the cloth item is also quite valuable to determine the ideal position of the gripper for a successful grasp. This information can be gained from lateral cameras (when the cloth is lying on a table), from stereovision or from range cameras. Some authors use this information alone, like the stereovision-based methods used by [30], who combine a minimum height threshold with a verticality criterion of the surface patch, or [31], who also aim at the highest point for grasping. However, it is much more common to use image segmentation together with depth evaluation, and examples can be found for any one of the following three segmentation types:

- a basic background segmentation is combined with stereo correspondence in [26];
- a simplified histogram-based Ohlander's method is used for relief evaluation together with height information gained from lateral cameras to determine the exact grasp location, inserting the fingers in the ditches at both sides of a crest, in [28] (this work is expanded towards the case of textured (non-solid) cloth in [29]);
- a Felzenswalb and Huttenlocher's graph-based segmentation is used in [27], which combines a minimum-spanning-tree algorithm with an adaptive estimate of the internal similarity of the clusters that it creates. Height is measured from stereo to obtain the item on top of the pile.

Alternatively, in a controlled scenario like an industrial environment, special illumination can be used to facilitate the process of stereo correspondence. This is achieved in

[32] by projecting a regular dot and mesh pattern on the items to be handled (towels in this case). Highest points of the crumpled towel, as well as corners are detected this way via shape extraction and contour tracing.

The arbitrary grasp point selected is generally the centroid or midpoint of the segmented region [26, 29]: the centroid is straightforward to compute and in general corresponds to a safe grasp point, with regards to the success of the grasp. Alternatively, the grasping point can be determined not as the centroid of the region but as the point that maximizes the distance to the region's border, through chamfering [27].

3.1.2. Wrinkle computation

An alternative to color-based binarization/segmentation plus height measure is to resort to some kind of wrinkledness measure, based on the evidence that wrinkled areas provide good candidates for successful grasps [33] (besides its use for cloth localization as mentioned above [15]). The depth information provided by Kinect cameras, specifically inclination and azimuth coordinates of the surface normals of a 3D point cloud corresponding to the cloth area in the image, is transformed to a wrinkledness measure by computing the entropy of the orientation histograms in a local area around each point. Points with highest entropy are chosen for grasping. The authors point at the convenience of combining this wrinkledness measure with a concavity measure, to avoid wrinkled but difficult to grasp areas.

Wrinkles at specific orientations are also aimed at for pick-up grasps by using the descriptors described in [34]. These 256-dimensional descriptors consist in 16 (4x4) spatial subdivisions with 2D angle histograms each (4x4 bins for azimuth and elevation angles). Each one of five different centroids (corresponding to the planar –non-wrinkled– case, lower-right, upper, lower-left and diagonal wrinkles, obtained after running k-means from a training database of descriptors) is used to rank the descriptors of a new perception. The grasping point is assigned to the descriptor closest to the selected centroid, with the corresponding gripper orientation. In this same reference, experiments with different perception-action couples are performed, using also the approach described above [33] together with different grasping strategies and scenarios with different number of garment pieces. Results are presented in tables that display the probabilities of grasping different number of parts: the measure of success could of course be the probability of grasping a single garment, if the pick-up is just a first step before other manipulations like folding, but it could also be the probability of grasping a higher number of parts if the objective is just a quick clean-up of the table.

Finally it should be stressed that both the highest and the most wrinkled area strategies for grasp point selection are considered as alternative actions within a broader POMDP-based action selection scheme for laundry separation in [35].

Figure 10 summarizes the process streams and alternatives for generic GPL reviewed in this Section.

3.2. Cloth features for grasping

It is often necessary to grasp a cloth item at specific, predefined locations. This is especially true when the cloth part has to be manipulated in a certain way, like in garment folding. To this end, these locations have to correspond to recognizable features, that can be identified by the vision system. Such features may be generic, like edges and

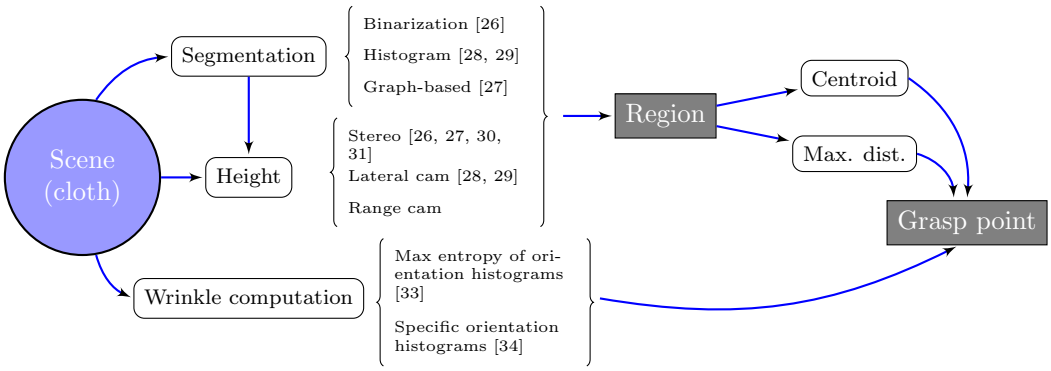


Figure 10: Different alternatives for generic GPL. Some references use only height (which justifies the arrow coming directly from the scene), whereas most works combine intensity image segmentation with height measure. The range camera is actually used in the wrinkle computation approaches, however it could constitute an alternative to stereo or lateral cameras for height computation.

corners, appearing in all kinds of cloth, or they can be garment-type specific, like collars, waistlines, cuffs, etc. In other cases the points to be grasped belong to a set of nodes of mesh models associated to different garment types. In such circumstances, rather than an identification of local shapes, what is needed is a global recognition process of the overall state of the cloth item (e.g., its shape when hanging from given points). Therefore, the latter family of specific grasp points identification is closely related to state recognition and classification processes (reviewed in Section 4). The different approaches taken for each one of these feature types are reviewed in the sections below.

The visual processing operates either on the cloth piece lying on a surface or grasped by the robot and hanging in the air. In the first case, generally the feature is intended to be identified from a single image, whereas feature identification on hanging cloth rather belongs to an active sensing strategy, where the cloth item is repeatedly regrasped and new images are taken and processed, aiming at a progressive reduction of the uncertainty about the grasped point. Common actions on the cloth in such active perception schemes to favor perception include –besides regrasping– shaking for disentangling [25] or for randomizing the item's configuration [26] and rotation at fixed intervals [25, 36, 26, 31]. Such actions can also affect the camera, not only for obtaining stereo from motion [37, 38], but just to obtain as much visual data as possible by moving a camera mounted on a robot up and down along the cloth [31] (together with rotation of the cloth, in this reference).

3.2.1. Simple features: edges and corners

These features are extracted from outline information, which in turn results either from a foreground-background segmentation or from edge detection methods. Corners are looked for in [22]: the outline information is encoded as a sequence of edges and corners, and a partial pattern matching procedure against the known unfolded fabric outline allows to estimate the location of the corner to be grasped for the following unfolding action of the robot. The first phase of the unfolding procedure described in [39] does also operate on the binarized or background segmented image of the garment item: the extreme edges in a set of eight discrete directions (easily to detect) are then chosen for grasping and pulling. More involved is the corner detection in the second phase, which uses depth

information to determine the *peak ridge* (binarized area of the cloth image whose depth is within the 10% of the depth of the highest point), the Harris corner detector for the corners at the border of the cloth image, discontinuity filters to detect sharp changes in depth, and continuity checks for the peak ridge and the peak corner locations, that allow to choose the latter as candidates for grasping. Corners as targets for unfolding are also aimed at in [40], who rely on the popular Canny edge detector to infer the location of corners from local maxima of contour curvature. The corresponding "unfolding axis" is also found as the edge opposite to the corner. Varying lighting conditions favor the occurrence of unwanted edges or of gaps on real edges, which in most cases can be handled by adjusting the threshold of the Canny detector. Stereoscopic imaging should enhance the overall performance, so the authors. The main contribution of this work is a systematic characterization of corners that may appear on a folded garment, as illustrated in Figure 11.

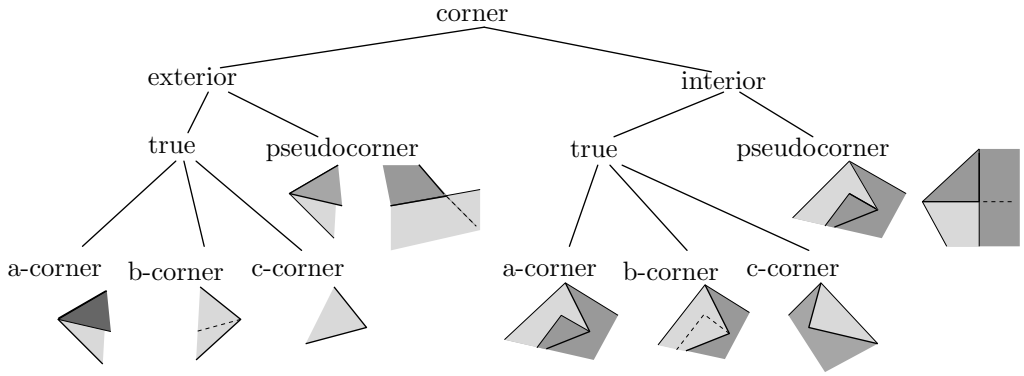


Figure 11: Taxonomy of corners on a folded part. The first criterion is obvious, exterior corners happen on the outer contour, whereas interior corners are inside the region corresponding to the cloth item. The distinction between true corners and pseudocorners lies on the considered edges of a triplet of coincident edges in that corner: the true edge is formed by the outer edges, whereas the pseudocorner by one outer and the inner edge. Pseudocorners do also arise when one edge crosses another one. Finally, a- and b-corners arise from foldings, with the folded part being above the garment and visible or below and invisible, respectively, and c-corners are the real corners on the garment when completely spread out. Note that a corner may be simultaneously both an a- and a b-corner. Adapted from [40].

All the preceding contributions assumed the cloth item to be lying on a surface like a table. However, corners can be also the perception goal when the piece of cloth is hanging from the robot's gripper. Regrasp-based solutions resort on combining simple perception procedures with basic manipulations in order to end up grasping the clothing item at the desired point. Usually the cloth part is assumed to be already grasped by the robot (which can be achieved, e.g., by the procedures described in the previous section) and hanging. One of the most obvious clues is to consider the lowermost point of the cloth region as a good candidate for regrasping [29, 25, 41]. In [29], resorting to the lowermost point (obtained from the cloth region outline) is complementary to two hemline detection procedures based on segmenting the image to determine the shadow region, and is applied when these fail. The application of the three procedures enables a bimanual grasping of the garment at two points on different hemlines. The lowermost point is likely to be

a corner of a garment part [25], and to correctly identify it both a shaking maneuver for disentangling the cloth part and a rotation at fixed intervals to determine the view where such point displays maximum curvature are performed. The action of grasping the lowermost point is executed repeatedly (alternating hands) for cloth part recognition, as explained in Section 4.

Still aiming at detecting a corner while holding the garment piece in the air, the approach followed in [37, 38] resorts on a special gripper design, with an integrated seam-following mechanism. As for the sensing aspect, besides fototransistor couples at the fingers, whose activation pattern varies according to whether the gripper is holding an edge or a corner, an external 8-bit grayscale CCD camera provides confirmation of corner detection, as well as 3D localization of the corner for grasping, by using motion stereo. The authors report success in corner detection even in absence of the external camera.

In [26] a stereo multi-view border classification algorithm, which discriminates actual garment borders from folds based on a curvature criterion, is combined with a RANSAC algorithm to detect grasp points located at the corners of a towel, exploiting a strategy of grasping two consecutive corners for folding. Several attempts to find a suitable grasp point by the corner detection algorithm are performed by rotating the cloth in front of the camera in small angle increments, preceded by simple shaking maneuvers to randomize the configuration of the towel.

Within a context of bimanual robotic unfolding of hanging garments, with repeated regrasps and matching to folded templates, the work in [42] aims at outline points, more specifically to points that bring the garment into a half-folded state. Once in such a state, the garment is matched against a polygon model of the folded garment, with predefined grasping point pairs, that are subsequently regrasped. The outline points are determined from depth by obtaining edges (Canny edge detector and Douglas-Peucker algorithm to separate edges of different orientations) and applying a set of geometric rules.

Table 1 summarizes the algorithmic features of the mentioned edge and corner detection references.

3.2.2. Specific features: garment elements

Garment features can, in isolation or in sets, constitute discrimination criteria for distinguishing different garment types (see Section 4.2 for this use). Some of them, like cuffs, elbow patches, collars or epaulettes are also appropriate as grasping points because they are recognizable, because of their graspability, and because once grasped and lifted (specially in the case of bimanual grasps) the resulting cloth state is suitable for subsequent manipulations. Representative for this approach is [43], where specific features like collars in polo shirts are aimed at as good grasping point candidates. Such components are identified, even if clothes are in a highly wrinkled state, by building up a Bag of Features based detector, combining appearance and 3D geometry features. Images are background segmented using a simple color threshold and then scanned by a sliding window with a linear classifier (logistic regression). Candidate windows are further ranked by using a more expensive but more reliable χ^2 Support Vector Machine (SVM) classifier, and finally the grasping point inside the best candidates is determined by using the wrinkledness measure mentioned above and described in [33]. Appearance-based filters were also used earlier in [44], to identify –by cross-correlation– potential grasping areas derived from user-defined samples and corresponding generically to edges and wrinkles or folds that

Reference	Input	Detection algorithm	Output
[22]	Binarized	Partial matching	Corners (contour)
[39]	Binarized	Extreme edges in discrete directions	Edges (contour)
[39]	Depth	Peak ridge detection	Peak corners (interior)
[40]	Grayscale	Canny edge detector	Corners (interior)
[29]	Grayscale	Histogram-based shadow segmentation	Hemline (interior)
[29]	Grayscale	Lowermost point	Corner (contour)
[25]	Grayscale (special illumination)	Maximum curvature by lowermost point	Corner (contour)
[32]	Projected pattern (special illumination)	3D shape extraction + Contour tracing + Surface curvature	Corner (contour)
[37, 38]	Grayscale + fototransistors	Activation patterns	Corner (contour)
[26]	Stereo multi-view	Curvature criterion + RANSAC	Corner (contour)
[42]	Depth (Xtion)	Canny + Douglas Peucker + geometric rules	Corner (contour)

Table 1: Comparative algorithmic highlights for edge and corner detection methods, both for lying and for hanging clothes.

the operator considered to be graspable. Without specifically addressing grasping issues, but expanding the set of garment parts to detect as suitable grasping candidates, the work in [45] proposes a learning and detection methodology that they apply at 11 such features in their experiments. Their detectors are appearance-based (SIFT features) but take also depth into account, as data are collected with a Kinect camera, by using the Geodesic-Depth Histogram (GDH), the Fast Point Feature Histogram (FPFH), the Heat Kernel Signature (HKS) and the Fast Integral Normal 3D (FINDDD) descriptors (consult the paper for brief explanations and references for each one of them). These are combined into a Bag of Visual Words descriptor (BoVW). As in the previous reference, logistic regression is combined with local χ^2 SVM for classification.

In [41], specific points like the shoulders of shirts and T-shirts or the corners of the waist of trousers and shorts are aimed as grasping points for unfolding the garment in the air. To this end, they also grasp first the lowermost point of the hanging cloth item, as this dramatically reduces the number of possible states of each garment type and thus paves the way for recognition and grasp point estimation. Once the test item is classified (see Section 4), the location of the key-points for grasping is estimated using Hough Forests, a similar formalism to random forests with extra information regarding the location of the grasp point. Using also an active sensing strategy based on a Partially Observable Markov Decision Process (POMDP), two trained Hough Forest for each garment type –one for the first grasping point and one for the second one– provide the probabilities of observing the grasping points at the different locations together with the probability of the grasping point being visible (i.e., not ocluded by other parts of the garment). The

POMDP formalism encodes the set of discrete hung-up states as well as the rotations and possible grasping actions for the locations of the corresponding grasp points.

A summary with some algorithmic highlights of the referenced contributions is shown in Table 2.

Ref.	Input	Image features	Learning/Detection algorithm	Detected features
[44]	Colour (on-board cam)	Appearance filters	Cross correlation	User defined edges and folds
[43]	Colour + Depth (Kinect)	BoF: Appearance (SIFT) + 3D geom	Logistic regression (global) + χ^2 SVM (local)	Collar polo shirts
[45]	Colour + Depth (Kinect)	BoVW: Appearance (SIFT) + Depth-based features	Logistic regression (global) + χ^2 SVM (local)	11 features (collars, hemlines, hips, sleeves)
[41]	Depth images	Hough forests	POMDP	Shoulder + waist corners

Table 2: Comparative algorithmic highlights for specific cloth features detection methods.

3.2.3. Node points of cloth models

When a mesh model of the handled clothes exists, specific nodes on this mesh can be chosen as grasp targets. This choice can be done online following some criterion (like in [46]) or just predefined (as in the works of Kita et col. or in [47]). And as in Section 3.2.1, the cloth item can be either lying on a surface (again [46]) or hanging from the gripper/s (as in the other references commented in this section). Clearly, such grasp points cannot be identified locally, but require to determine the state of the cloth item, i.e., to establish the correspondence between the 3D model of the cloth and the view displayed in the image. Thus, this section is closely related to Section 4 below, but here the accent is put on the localization of grasp points.

In the context of a cloth folding application on a table, the skeletal model (a skeletal mesh with additional parameterized landmark nodes) in [46] is fit to the contour of the observed image, and the landmark points of the model are relocated to the nearest neighbor on the contour. The points to be grasped among these are selected as shown in Figure 12. It should be noted that the method is not only applied to the original spread-out shape, but also to the different folded states.

The two-phase approach in [23] is somehow inbetween this contribution and the works presented below, as the cloth item is hanging in the first *disambiguation* phase, while repeatedly laid on a table and regrasped during the second *reconfiguration* phase. The first phase aims at classification and state estimation using an active sensing Hidden Markov Model-based strategy while repeatedly grasping the lowermost point, and is explained in Section 4.1.2. In the second reconfiguration phase (which is executed unless the current state after disambiguation is not already the desired final state), the actions to be planned with consist in repeatedly laying the garment on a surface, opening both grippers and pick up the item at a new pair of points. To determine the sequence of such actions, a *graspability* graph is constructed with edges between the grasp point pairs such that in the cloth configuration determined by one of these pairs, once laid on the table (assuming

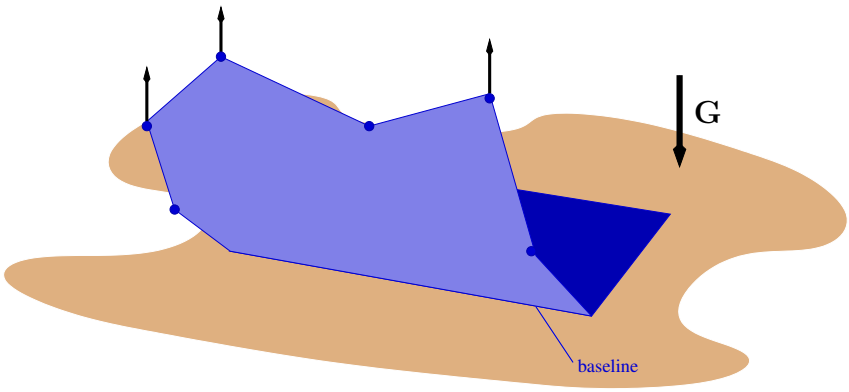


Figure 12: The points to be grasped are determined by the criterion of avoiding the deformation of the part of the cloth which is lifted, hanging vertically from the grippers: all the convex corners not on the baseline (i.e., not touching the ground) at which the negative gravity vector does not point inside the cloth polygon have to be grasped. These points are marked with an upward arrow in the figure.

this is done without changing the hanging configuration) the other one is *graspable*. A point is assumed to be graspable if no other parts of the cloth can be simultaneously grasped when the grippers close: the Euclidean distance of the checked point to all the other points in the mesh separated by a large geodesic distance should be larger than a minimum threshold. Once constructed, a Dijkstra graph search from the initial pair to the goal pair of held points provides the path on the graspability graph that corresponds to the seeked sequence of actions.

As for hanging clothes, in [2, 36, 48, 49, 30] grasp points are previously set on specific locations of garment models, and the image processing part is endowed with the task of matching the observed cloth with the possible shapes of the deformed mesh model, so that the point to be grasped is also located in the observed image. In [2], separated 2D visual information is used for matching right and left images, and an additional correspondence process provided the necessary 3D information to locate this point in space. State recognition and grasp point localization could be performed simultaneously in later settings [48, 49, 30], as their new trinocular stereo vision system allowed to attempt directly the matching of 3D observations with the 3D deformable triangular mesh models they use. Once identified, the information about the localization of the (predefined) grasping point is readily available, and the required gripper orientation can be easily computed [48].

The features to be grasped can be any user-defined pair of points on a garment. The discretization provided by a mesh model facilitates the choice of these points, as well as the regrasping strategy that leads towards the desired grasping state [47]. After pick-up and pose estimation (see Section 4 for details on their previous works on this issue), the regrasping point is found by optimizing an objective function that measures the closeness to the desired grasping points (scaled by the predicted probability of the current grasped points being accurate). Such metric is evaluated over the whole surface of the model, and the optimal grasping point in the real image is determined by registration. The method to find point correspondences between the source (model) mesh and the target mesh obtained from 3D scans of the grasped garment proceeds in three steps, that include scaling, rigid transformation (by an iterative closest point procedure), and a non-rigid registration.

The latter process balances an energy term penalizing discrepancies between source and target meshes and another term relative to the required deformations of the source mesh. This gives a *global localization* of the regrasping point, and the final gripper position and orientation are obtained by local refinement using a 1D blob curvature detection algorithm with the information provided by an IR range sensor. The authors report 1-2 necessary regrasps for successful unfolding in most cases.

The complexity of an image processing–intensive state-recognition method is avoided in [31] by using coded fiducial markers over the whole T-shirt manipulated by a PR2 robot. This shortcut is taken because the accent is put on grasp planning rather than on the visual identification of the different parts of the garment. The goal of the manipulation is to end up with the two arms of the robot holding the shoulders of the T-shirt, and to this end they follow a multiple regrasp strategy. As each marker has a different code, simple inspection allows to quickly bring the observed markers into correspondence with a mesh model of the cloth. This model has been generated before and automatically from photographs of the extended T-shirt. Despite using this sidestep, this work deserves to be explained here for two reasons: first, it makes a simple but smart use of these markers to identify the grasped point as the one minimizing the difference between the geodesic distance on the cloth surface and the Euclidean distance to the other points (this holds only for the grasped point, if the hanging cloth can be thought of as a cone grasped by its apex). The mesh model allows for a straightforward computation of geodesic distances. Second, an active perception strategy is followed for identifying the next grasp point. To this end, as many as possible point clouds, position of folds and positions of markers have to be collected, by rotating the cloth in front of the stereo camera and by moving the monocular camera mounted on the forearm of the free robot up and down. Fold lines are detected by filtering shadow gradients and applying the Hough transform. The next grasp point is obtained by a greedy policy: the candidate is the point that brings the next grasp closer to the target configuration. With these elements, the grasp itself is generated by identifying the fold lines next to the candidate point and through a collision detection procedure between the boxes representing the fingers and palm of the gripper and the point cloud next to the candidate. The choice among the possible grasps generated in this way relies on a score function that evaluates their likelihood of success. This function has been previously learned: features like the number of cloud points between the gripper's fingers or the average surface normal of cloud point close to the fingertips are recorded during training together with a success/failure label of the executed grasp. A SVM with a Gaussian kernel function is used for learning during the training phase. Training is done automatically, by consecutively grasping random points and releasing the previously holding gripper if the grasp is successful.

Table 3 provides a quick overview on the described algorithms.

4. Cloth state identification and classification

Closely related to grasp point selection (in fact, some of the references are the same), the tasks of classifying a specific garment within a set of cloth categories (like "pants", "blouse" or "towel", each author specifies her own set of cloth types) and to determine in which state it is (basically how/where it is grasped) clearly are the outcome of an image recognition process. Most of the references assume the cloth part is hanging, grasped at

Reference	Input	Model-observation correspondence	Node selection
[46]	Colour	Contour fitting to skeletal model	Landmark nodes of model
[23]	Colour	Contour fitting to mesh model (DTW)	Node pair in graspability graph
[2]	Colour stereo	2D matching of left and right image + Correspondence of nodes	Predefined
[36, 48, 49, 30]	Trinocular stereovision	3D matching of triangular mesh	Predefined
[47]	3Dscans	3D fitting with scaling rigid transformation, non-rigid registration	Optimization closeness meassure to desired gp
[31]	Fiducial markers	Marker-node matching	Greedy policy: next grasp closer to target configuration

Table 3: Comparative algorithmic highlights of node point detection methods.

one or two sufficiently distant points, but there are also works where classification is done while the garment lies on the table, spread out or even crumpled (see Fig. 13).

Most of the works referenced below perform both tasks, classification and state recognition. However, there are also some works that assume the cloth type to be known beforehand and therefore concentrate on state recognition [2, 36, 48, 49, 30], and others that only try to classify the garment piece. As for the latter, restricting to classification is either due to the application at hand (e.g. laundry classification in [27, 50, 51]) or because the cloth item is assumed to be already in a standard configuration (approximately spread out on a table in [4]). Categorization of classification and state recognition algorithms can be done by distinguishing whether the employed method is based on the overall shape of the displayed item or relies on the recognition of specific local features.

4.1. Shape recognition methods

These methods compare the shape of a cloth piece in the current image either with synthetic (i.e., computer-generated) cloth images [24, 2, 36, 48, 49, 30, 23] or with images of real clothes [25, 27, 4, 52, 41, 53, 5] in a database. Shape comparison can be performed either by comparing some characteristic measures of the shapes or by matching of the contours.

4.1.1. Geometric shape features comparison

The works base their approach on comparing geometric values that are easy to extract, like the area of the binarized isolated image of the cloth. This idea is perfectly illustrated in the two-stage approach in [24], which works with the silhouette of the cloth grasped at two distant points, and stretched by pulling the grasp points away form each other. The convex closure of this silhouette allows to perform a first rough classification into three types just by counting the number of convex absences (regions on the image not occupied by the cloth but inside its convex closure). These regions have to be larger than a certain threshold, and the region corresponding to the slack on top, between the grasped

Figure 13: Different appearances of a garment in which classification and/or state recognition algorithms have been tested. Hanging states may, in turn, be just an arbitrary configuration of the garment held at any point, or belong to a finite set of discrete configurations corresponding to predefined grasp points.

points, is excluded. This first classification does still not determine the garment type the item belongs to, but this –together with one of the predefined states of a discrete set the cloth may be in– is the output of a second refinement phase. This refinement requires the measurement of widths and heights of the stick-out regions and the non stick out region, as well as measures of the slack region (see Figure 14), and ratios among these measurements (for normalization) are computed. The decision process is a kind of flow-chart where such ratios, plus the number of detected stick-out regions, and the proportion of the areas of the stick-out and non-stick-out regions, are compared to the models' ones. A total of 18 stored silhouettes of towels (7), pants (3) and shirts (8) are used for categorization of both cloth type and its hung-up state. These states are discretized by considering only a finite set of grasp point localizations along the outline of the cloth. The authors suggest to use virtual physical-based models, simulate their grasping as described, and obtaining virtual binarized images, "aspect models" to be compared with the actual image.

Using real acquired and stored images instead of model-based ones, the interactive regrasping and sensing procedure described in [27] stands out for the randomness of the picked-up point, both in the database and in the test images. The clothing item is just grasped as described in Section 3.1, lifted, a side-camera takes two images with a rotation of $\frac{\pi}{2}$ which are binarized and compared with the database, before dropping the item and repeating the procedure up to ten times. A nearest-neighbor algorithm is used for matching all the test images with the labelled database images, using as features the absolute difference in area of the two silhouettes, the absolute difference in eccentricity, the Hausdorff distance between their edges, and the Hausdorff distance between the Canny edges of the original grayscale images, yielding the best results for combinations including the first two features.

Figure 14 summarizes some of the geometric features used in the two mentioned works.

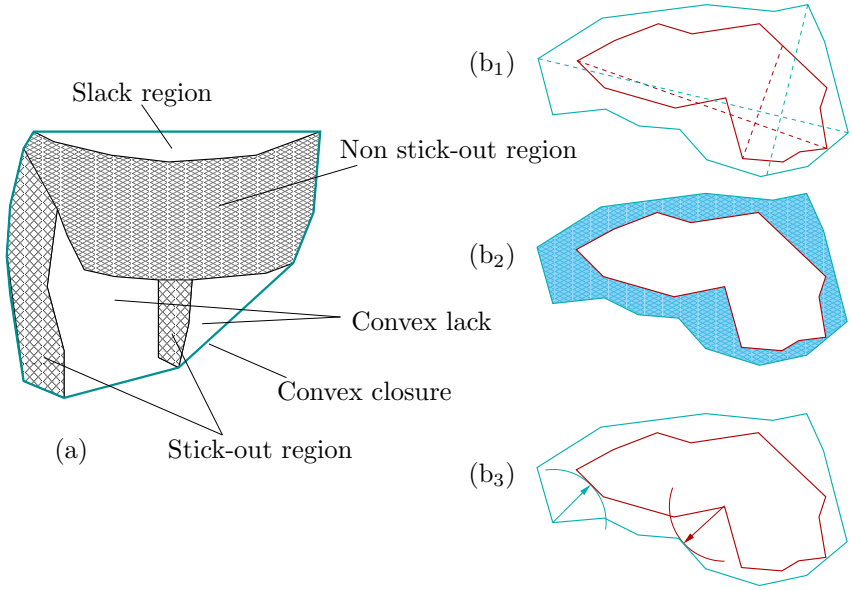


Figure 14: (a) Geometric features whose measurements and their ratios are used in [24]. (b) Some of the geometric features used in [27]: (b_1) Eccentricity difference, where the eccentricity is measured in terms of the ratio between the maximum chord of the shape and the maximum chord which is perpendicular to the first; (b_2) Difference in area (shaded region); (b_3) Hausdorff distance between contours.

Table 4 summarizes the geometric measures and matching procedures used by the two described contributions.

Reference	Input	Geometric measurements	Matching with model
[24]	Silhouette	Convex absences + width/height stick out regions	Flow chart comparison (one shot)
[27]	Silhouette	Absolute area difference + Eccentricity difference + Hausdorff difference (binarized edges, Canny edges)	Nearest neighbor (multiple regraspings)

Table 4: Comparative algorithmic highlights of geometric features based methods for classification and state estimation.

4.1.2. Contour matching

Strictly speaking, this approach could be included in the previous section, where, in fact, already a contour-coincidence measure expressed by the Hausdorff distance between the silhouette edges was mentioned. However, while in that reference contour comparison was just one more measure, here –even if not exclusive in the case of [23]– contour matching is really the discriminating feature among garment types and states. Contour matching can be approached either by applying some basic transformations on the models' contours and then resorting to a simple coincidence metric, or by using a more sophisticated procedure. These two options are represented by the works explained below.

In [54, 2] basic transformations are performed on the models in order to attain the highest degree of coincidence. Simulation is based on 3D quadrangular mass-spring models (including diagonal springs), held and hanging at any one of the nodes, and are projected on a vertical plane. This projection is what can be compared to the real 2D garment image held by the robot, obviously also with a one-hand grasp. No classification is done as the shapes of the garment parts (pullovers, trousers, ...) and their approximate size are assumed to be known in advance. The basic transformations for appearance matching between the contours include a vertical displacement of the projected mesh model, as well as a width normalization (shrinking or extending the model appearance horizontally). Afterwards, a simple overlap ratio is used (the sum of the ratio of the overlapped area to the model appearance area plus the ratio of the overlapped area to the observed area). The authors report that only specific colors and textures of the cloth parts guarantee robustness in the observed cloth region extraction process. A region overlap criterion does also lie in the basis of their later 3D mesh matching works [36, 48, 49, 30] (see Section 4.1.3).

Simple overlap may perform poorly, as illustrated in [23] and in Figure 15 below, and thus other more elaborated contour matching algorithms have been developed, based on nearest neighbor distance [4] or on dynamic programming [23, 53, 5]. In the first case, the state of the cloth is assumed to be known beforehand, as all the garments are presented loosely spread out on a table. The image of the cloth and the set of available category models (their *parameterized* models, as mentioned in the Introduction) are presented to an energy optimization algorithm, where the energy function makes use of the average nearest-neighbor distance between the generated contour from the model parameters and the contour of the image. This distance constitutes a measure of overall model fit, and the image is classified as belonging to the category exhibiting the lowest energy. The authors claim to have considered also dynamic time warping (DTW), observing only little improvement, whereas nearest neighbor is much easier to compute. Nonetheless, a DTW algorithm for contour matching is at the heart of the Hidden Markov Model (HMM) based active perception strategy in [23] (explained below in Section 4.1.4). A more accurate contour alignment procedure may be justified in this case as not only the garment type but also its hung-up state (grasped at two mesh points) have to be determined. The contours of stored images of mesh models in minimum-energy configurations when held at given pairs of nodes are used for matching. The DTW algorithm provides a similarity measure for two sequences (the contours in this case). Here, pixel coordinates and first and second derivatives with respect to arc lengths are the chosen features for alignment (the derivatives stand for corners and similar features). Other features that can be chosen for sequential alignment are just the vertices of the contours, as done in [53, 5] (in [5] additionally the fitting of relative line segment lengths is included). Their dynamic programming algorithm attains polynomial complexity, and efficiency is further enhanced by the fact that matching is done between a *simplified* polygonal contour of the test image and the polygonal models of the garments. This simplified polygonal contour of the approximately spread out garment on the table is obtained in three steps: background segmentation, contour computation by the application of Moore's algorithm, and polygonal simplification of the contour by another dynamic programming algorithm that minimizes the overall distance to the original contour. This simplified polygonal contour

has typically much less vertices than the model polygonal contour. Vertex correspondence of the simplified polygonal contour vertices means that their angles have to be the best adjusting to the normal distribution of angles associated to polygonal model vertices, whereas the vertices inbetween are just "matched" against a generic segment (i.e., have an inner angle of π). These correspondences are illustrated in Figure 15.

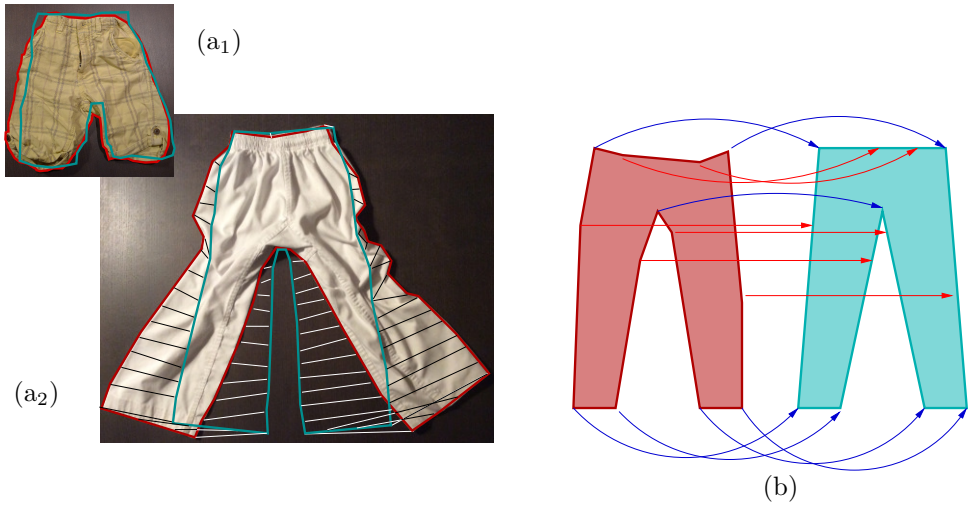


Figure 15: While in some cases a simple overlap ratio may suffice for assessing contour coincidence (a_1), in others a great discrepancy may exist (a_2). In such cases, the DTW algorithm provides a reasonable contour correspondence measure [23]. (b) Vertex-vertex (in blue) and vertex-edge correspondences between the two polygonal contours corresponding to the model and to the test image (adapted from [5] to the pants example).

Table 5 summarizes the algorithmic differences for contour matching of the referenced works.

Reference	Input	Model	Matching with model
[54, 2]	Colour stereo	3D quadrangular mass-spring mesh, vertical projection	Simple overlap ratio
[4]	Colour	Parameterized model	Energy optimization based on average nearest neighbor distance
[23]	Colour	Pixel coordinates + 1st and 2nd derivatives w.r.t. arc lengths	DTW of contours
[53, 5]	Colour	Vertices of contour	Dynamic programming

Table 5: Comparative algorithmic highlights of contour matching based methods for classification and state estimation.

4.1.3. 3D matching

Stereo vision and depth imaging allow to obtain volumetric information (e.g., in the form of a spatial cloud of points) of a cloth item, which can be directly compared to the 3D mesh models representing the different garment types. Alternatively, the classification can be based on volumetric information obtained by training on real depth images. Again,

as in the previous approaches, some kind of discretization is mandatory to cope with the infinite configuration space of deformations, which is also attained by grasping at a mesh node or resorting to a lowermost point grasping strategy. The first group of works discussed below corresponds to a somehow intermediate approach between 2D and 3D, as it uses the planar projections of 3D meshes for matching. The other contributions deal directly with volumetric information, either via SIFT features of depth images [55], binary cell occupancy vectors [56, 47], or Random Forests encoding depth differences between pixels and mean curvatures [41].

Like in their earlier works mentioned in the previous section, no classification is done in [36, 48, 49, 30]. Here 3D triangular meshes are used as models, and a trinocular stereo vision system allows to perform spatial matchings to identify the state. A key procedure, detailed in [36], is to allow the stored states to deform gradually to some extent until matching the observations, starting at the fixed grasped triangular patches (a planar grasp is assumed, which ensures the known localization and orientation of the grasped patch). These deformations are computed from physical analogues of forces, including the internal shape preserving forces that keep the distance between neighboring nodes and nodes sharing common neighbors (standing for elasticity and flexural rigidity of the cloth, respectively), as well as the external forces of gravity and of attraction to the closest observed points (only if the nodes belong to a certain list of "patches to attract"). An overlap criterion which uses the ratios of the coincident region to the observed and the visible model surfaces is used to decide to which state the observation belongs to.

In [55] the two phases of categorization and pose estimation of the input test images are addressed by means of a two-layer classifier. The garment-type identification layer uses an accurate Radial Basis Kernel function (RBF) SVM, whereas the grasped point-based state estimation layer uses a faster linear kernel SVM (combined with spatial max-pooling), as there are much more categories in this second phase. Again, this classifier is built up (trained) by using synthetic images provided by a physical simulator of mesh models. SIFT features are obtained from depth images collected by 90 virtual cameras distributed on a geodesic dome around the grasped garment simulation (hanging from one of the 20–50 predefined grasp points). From these features, a codebook for recognition is built up by sparse coding and dictionary learning. Experiments using virtual test images as well as real depth images of quite similar clothes obtained with a Kinect sensor are performed, with very good results. As for pose estimation, this work is enhanced shortly afterwards in [56] with a faster (real-time) and more accurate algorithm, where testing is performed on the images provided by a Kinect sensor from garments held by a Baxter robot. The set of grasp points on the virtual garment is obtained by uniform sampling on its 2D UV map and mapping them back on the 3D model. Simulation is similar to their previous work [55] but now the output are mesh models instead of rendered depth images, which have to be matched against a reconstructed model from the Kinect sensor. To perform this matching, binary feature vectors are used that encode volumetric information of the hanging garment, more precisely occupancy information attached to the discretization of a cylindrical bounding volume (using polar coordinates at each of the circular slices along the cylinder's height). The matching of two shapes requires to obtain the rotation that minimizes the Hamming distance between their feature vectors (i.e., their lowest bit-to-bit XOR-ing). With this optimal rotation obtained for all the models in the database

against the query, the nearest neighbor is found based on a weighted Hamming distance. These weights are previously learned from calibration data, as an optimization problem of minimizing the empirical error with a large-margin regularizer. Evaluation of the accuracy in the experimental stage is done using the geodesic distance between the predicted grasp point and the ground truth. It should be noted that in this work pose is defined by one grasping point, this is expanded to two held points in more recent work [47].

In contrast to these works, classification and state recognition in [41, 57] and in [58] use only real stored images. The work in [41] is based on an active recognition strategy implementing a POMDP framework, as described in the next section. The authors exploit the fact that for garments like shirts and trousers held by one gripper anywhere, there is only one possible lowermost point (not counting symmetries), located at the cuff and the heel respectively, whereas shorts and T-shirts display two such points. Considering these four garment types, a total of six classes arise when the clothing item is grasped at a lowest point. Random Forests are trained on depth images of such garments and graspings, associating binary tests at the nodes consisting in depth differences between pixels and mean curvatures. This Random Forest classifier provides the observations (probabilities of belonging to each class) to the POMDP strategy. This method is further enhanced in [57] by using Active Random Forests, that allow to perform classification, grasp point detection and pose estimation within the same tree structure: applying a hierarchical coarse to fine quality function for node splitting, the upper parts of the tree are responsible for classification whereas the lowest parts perform the other two duties. The second work [58] goes to the root of the probabilistic nature of the categorization problem and implements classification as a Gaussian Process (GP) strategy. RGB data are only used for isolating a garment of a pile, or for background segmentation, whereas classification bases exclusively on depth. The same high-resolution setting of [19] and described in Section 2 provides the required accurate depth data. The features computed from the depth-map and used in the multi-class GP classification procedure (using the Laplace approximation for posterior inference and optimizing hyper-parameters through marginal likelihood maximization) include the Shape Index histogram (SI) (also used in [19]), Topology Spatial Distance (TSD) and Multi-Scale Local Binary Patterns (LBP). The authors show how the classification accuracy is directly related to the confidence of the prediction, determined as the conditional probability of the testing sample given training examples. The experiments demonstrate that proposed features are far more reliable than other descriptors like the BoF used in FINDDD or the Volumetric Descriptor, and that the method is comparable to state-of-the-art references in a single-shot procedure, while it outperforms them when GP is used within an interactive perception strategy, as described in Section 4.1.4.

The pros and cons of generating real and synthetic data emerge in a learning context that is somehow extreme due to the large required size of the involved datasets, namely deep convolutional neural networks (CNN). This is the approach taken in [59] for classification and pose estimation of (single point grasped) hanging clothes. Gathering the real images is too much time-consuming, not only because of the grasping and lifting, but also due to the required slow rotation of the garment before the Xtion camera while taking pictures. Through simulation (Blender 2.6.2), on the other hand, a large set of synthetic depth images can be generated. Both types of data have been used in this work within

a two-layered deep CNNs structure, where the first layer stands for classification and the second for state recognition (again as one of the discrete points the grasped garment is hanging from). The two CNNs have the same depth and topology, but the size is considerably larger in the second one as the number of output classes is also higher in pose estimation. The authors have further investigated the cross-domain learning possibilities by testing the real data in the synthetic-trained classification CNN, yielding a classification rate that is not above chance. The real dataset has been further used for comparison with other SVM and random forests-based approaches. This very same classifier is later used by the same authors plus collaborators in a multisensorial framework for garment type, fabric pattern and material classification [60]. RGB-D data (i.e., RGB plus depth) are the input to the fabric pattern recognition and, together with tactile and photometric stereo sensing, also material recognition, by using Random Forests classifiers in the two cases, although on features arising from Gabor filtering in the pattern recognition case and HOG in the case of material recognition. As for the garment type recognition, it bases exclusively on RGB-D data, and the classification decision bases on the output of two classifiers: the already mentioned CNN operating on the raw sensor data, and again a Random Forest on HOG features.

Table 6 stresses some algorithmic features of the presented 3D matching methods.

Ref.	Input	Model features for comparison	Model adjustment/ Model encoding	Matching
[36, 48, 49, 30]	Trinocular stereo	Planar projection of 3D triangular mesh model	Gradual deformation of model	Overlap criterion for SE
[55]	Kinect sensor	SIFT features of 3D mesh model/real depth images	Codebook (sparse coding + dictionary learning)	RBF SVM (class) + linear SVM (SE)
[56, 47]	Kinect sensor	3D mesh models	Occupancy of discretized cylindrical volume	Rotation, nearest neighbor based on Hamming distance
[41, 57]	Depth	Depth images of set of garments held at discrete nodes	(Active*) Random forests	POMDP strategy
[58]	High resolution depth	SI + TSD + LBP		Multi-class GP
[59]	RGB-D (Xtion)	Synthetic depth images	Deep CNN	CNN classification
[60]	RGB-D (Xtion) + tactile + photometric	Gabor filters + HOG	CNNs + Random forests	CNN on raw data + Random forest on HOG

Table 6: Comparative algorithmic highlights of 3D matching based methods for classification and state estimation. The term "Model" in the heading of this table is used in a broad sense, meaning not only synthetic but also real images used for comparison with the observed data. (*) in [57]

.

4.1.4. Active perception

Besides the primal goal of manipulation of a cloth piece (folding/unfolding, separating, assisting people, etc.), manipulative actions may enhance the perceptual objective of

categorization and state identification. Common actions (some of them already mentioned in Section 3.2), include rotating the cloth item while grasped and hanging, [25, 36, 49, 26, 31, 41, 59, 60], shaking [25, 26], or using a second hand to spread by pushing the cloth part in front of the vision system [49]. Among the rotation-based works, [41] deserves to be stressed as the rotations and the image taking are inscribed within a POMDP framework, similar to the one described in Section 3.2.2 for grasp point localization. Here, the overall shape is analyzed instead of aiming at specific features like shoulders or cuffs as grasp points. As said above, observations are provided by the Random forest classifier. This POMDP is specified further by the states, that correspond to the six classes (garment type + grasped point), the actions including one discrete rotation and the six recognition decisions, the transition probabilities that are 1 for the rotation action leading to the same state (zero otherwise) and equal to the initial belief state for the other actions, and finally the rewards, that penalize rotations and incorrect state recognitions. As for other active sensing solutions, we have already pointed at the dropping and random regrasping strategy in [27] (Section 4.1.1), which does also provide different views of the same cloth item, as fixed angle rotation around a vertical axis does. Special attention has to be paid to the processes where actions lead to intermediate recognizable or at least disambiguating states, as shown next.

Repeated lowermost point regrasping is such an active sensing strategy. This bimanual handling procedure consists in grasping the point of the cloth located at the lowest height while the other hand holds the cloth item. While the first grasp is undetermined (the cloth item can have been grasped at any point), its lowermost point when hanging can only belong to a certain set of points (e.g., one of the four corners in the case of a towel). Regrasping at this point (while releasing the previously held point) further reduces the set of possible lowest points (following with the case of the towel, the state becomes determined, up to symmetries). Normally, after a few regrasps this strategy ends up alternating between a couple of states. The last action consists in grasping the lowermost point without releasing the held one and extending the cloth holding the two grasped points at the same height. This strategy is followed in [25] and in [23] on similar sets of clothing articles (short and long-sleeved shirts, trousers, towels [25, 23], underpants, brassieres, and handkerchieves [25], skirts and a couple of infant clothing items [23]), with some differences in their approaches. A database of real template images is used in [25] whereas [23] resorts to a cloth simulator to provide the models to compare with. More significative is that while [25] base their recognition decision on the images taken during the last extension phase (uncertainty is reduced by not only considering the final image but also intermediate images at fixed intervals of this extension or straightening of the cloth), a probabilistic approach is followed in [23], who ground their decision process on a HMM which tracks the garment's configuration throughout this sequence of manipulations and observations. In this model, the hidden state is defined by the garment's identity or type, and by its grasped state, and the observations consist in the height of the hanging item (during regrasping), and its contour (at the end of the final extension phase). The transition model of the HMM gives the probabilities of each mesh node being the lowermost one and thus being grasped by the free gripper, reflecting the rapidly converging behavior of the lowermost point grasping strategy to a few alternating states (for each category). Finally, in [25] form recognition is performed by computing degree of

similarity coefficients based on covariances of pixel intensity of the candidate and stored images (after foreground-background segmentation and size normalization), whereas [23] perform contour matching as explained in Section 4.1.2).

The foregoing actions are applied on hanging clothes, but there are also perception-enhancing actions applied on clothes lying on a surface, at least in their initial state. This is the case of [58], where the action *Grasp-Shake*, followed by dropping the garment back on the table, favors the spreading-out of the clothing item, whereas the action *Grasp-Flip* aims at displaying previously hidden parts of the garment. The first action is available if the height of the garment exceeds 5cm (as the grasp is performed on a wrinkle) and the second if the thickness of the edge to be grasped is less than 5cm (wrinkle and edge grasps are explained in detail in their previous work [19]). The action to be performed is randomly chosen if both of them are available. Adopting such an active perception strategy definitely leads towards outperformance in garment classification w.r.t. single-shot strategies, as shown by their experiments.

Table 7 compares the actions and matching algorithms used in the most significative presented active sensing strategies.

Ref.	Actions	Matching algorithms
[41]	1st lowermost point + discrete rotations	POMDP strategy
[27]	Dropping + random regrasping	Nearest neighbor of all images
[25]	Lowermost point regrasp	Template matching of images of last extension phase
[23]	Lowermost point regrasp	HMM of different grasp/hung state images

Table 7: Active sensing strategies compared as for the perception enhancing actions and the matching algorithms, with specification of the set of images considered in the matching process.

4.2. Feature recognition methods

Garment types may be distinguished by sets of category-specific features, that range from fabric type [50] up to garment elements like sock heels and toes [52] (although here they are used for state identification), buttons, pockets and the like [51]. In most cases recognition will be more reliable if such features are not used in isolation, and some features should have a larger identification weight than others: for example, pockets may be useful for distinguishing shirts from other sleeved garments like pullovers, but clearly the frontal buttoned opening has a greater discriminating power, as many shirts do not have any pockets at all. Continuing with the shirts example, the granularity of the classification output should also be established beforehand, as it could just be the category "shirts", or more refined into "no-pocket shirt", "one-pocket shirt" or"two-pocket shirt" (plus all the combinations of possible collar and cuff types...). While most of the shape-based recognition methods require the cloth item to be hanging, or, if lying on a surface, to be more or less spread out, the contributions in feature-based recognition do only require one or more features to be visible, and the garments can even be in a crumpled state [50, 51].

In [50] the Gabor-filtering strategy used in [15, 17] for cloth detection (see Section 2) is extended to garment classification, by deriving a set of features based on detected wrinkles, cloth overlaps, scale space extrema and contour. As for the latter, this feature should not

be confused with the contour matching methods explained in Section 4.1.2 (here the cloth is not laid out or extended, it is crumpled), but it is just a distance histogram, obtained from the polar coordinates of the contour measured at the circumcenter of the cloth region and discretizing angles. Supervised multi-class SVM (with a radial basis kernel function) is used for learning and classification. The success of their classification method is due in large extent to the distinctive identification of fabrics type through the aspect of the generated wrinkles when the garment is randomly thrown over the table.

Also the work in [51] is quite feature-intensive and follows a multi-level approach to the classification problem: low-level features, both global (Color Histogram, Histogram of Line Lengths, Table Point Feature Histogram, boundary) and local (Scale-Invariant Feature Transform (SIFT), and Fast Point Feature Histogram (FPFH)) are used to determine and classify the mid-level *Characteristics* (i.e., cloth features like buttons, pockets, hemlines...). A SVM-based algorithm is employed to this end. Still in the mid-level, Selection masks are determined from the database and used to assign the set of characteristics of the candidate cloth to the corresponding high-level class. The use of multiple features extracted from 2D and 3D imaging collected together in a large feature vector resembles the work in [43] mentioned in Section 3.2.2, but here it is used for classification purposes.

For a particular case in which the category is already known (socks) but state recognition has to be performed [52] do also resort to a combination of 2D texture and shape-based features. The canonical modes the sock can display include sideways, heel up, heel down, and bunched states, all in the rightside-out or inside-out variants. The authors use two texture-based (MR8 filter bank and Local Binary Patterns (LBP)) and one shape-based feature (Histogram of Oriented Gradients (HOG)), and find out that the combination that performs best for inside-out vs. rightside-out classification is LBP plus HOG, while training the SVM classifier with a χ^2 kernel. As for state recognition, local detectors are trained with the appearance features to respond to image patches corresponding to the opening of the sock, the heel or the toe. Landmarks are placed on the centers of these patches and then compared with the parameterized models of each sock configuration. Finally, matching of pairs of socks is also accomplished by further using additional cues related to size and color.

Table 8 displays the main algorithmic traits of the referenced feature-based methods.

Ref.	Features for comparison	Classification procedures	Output
[50]	Wrinkles, cloth overlaps, scale space extrema, contour	Radial basis SVM	Fabric type
[51]	Low level: global (hist, cont) local (SIFT, FDFH) Midlevel: buttons, pockets,...	SVM (low $\longrightarrow$ mid) + Selection masks (mid $\longrightarrow$ high)	Garment type
[52]	Low level: MR8 + LBP + HOG Midlevel: sock opening, heel, toe	χ^2 kernel SVM + Comparison with parameterized model	Socks state estimation

Table 8: Comparative algorithmic highlights of feature-based classification and state estimation.

5. Conclusions

Cloth challenges the current computer vision state of the art. Fundamental processes as registration, parsing, recognition or tracking have to face the inherent problems of high

variability in appearance and pose, as well as self-occlusions. The standpoint from which this survey has been written is that of robotic manipulation of cloth. Thus, the focus has been set on the localization of possible candidate points for grasping a given cloth item, to classify a garment, or to determine the various states a piece of cloth may be in. These basic perceptual goals, together with cloth detection (as a material), are quite generic, and other more specific application-oriented ones, like following the progress of a given manipulative action on a cloth part, have been deliberately left out.

The three tasks –grasp point localization, classification and state recognition– are clearly related to one another, if not intertwined. Some grasp point localization methods require to assess the state the cloth item is in (Section 3.2.3), and –conversely– some classification and state recognition methods require the cloth to be grasped at specific points or rely on an active regrasping and perception scheme (Section 4.1.4). Despite classification and state recognition have nominally different goals, in practice they have to be simultaneously satisfied when an unknown shape (i.e., when neither the cloth type nor its state) is presented to the vision system. Classification can be certainly disconnected from state estimation, e.g. in some laundry separation applications, but in most cloth handling applications, even in type-dependent sequences of actions like garment folding, the initial state has to be known. Conversely, state estimation can also be a standalone task in the cases where a single type of cloth items (e.g. towels in [32] or in [26]) are dealt with. The connection between grasp point selection and state estimation is quite tight in the case where the state of the cloth item is determined by the point (one or two of a discrete set of nodes) where it is grasped and hanging. Furthermore, some active sensing techniques alternate between state recognition and grasp point selection, in a cycle that includes repeated regrasping, as seen in Section 4.1.4. Active sensing techniques are very effective in reducing ambiguities and uncertainty about the state of the cloth item or the point to be grasped. They have been mainly applied on hanging clothes (with the exception of the dropping and regrasping strategy of [27]). Less explored is the case of applying such techniques on lying clothes, where actions to enhance perception may include dragging and unfolding. The work in [3] can be quite inspiring in this sense.

Generic GPL has the difficulty of determining the 3D coordinates of an arbitrary point within a region of the image corresponding to the cloth item to be grasped. Furthermore, such point should be graspable. The consequences of an incorrect GPL are either grasping simultaneously more than one cloth item (from a laundry pile, for example), or grasping nothing at all. Some researchers consider that it is enough to segment the image and measuring the heights of the cloth within the segmented patch: the highest point is a good candidate for grasping. The complexity of this method is related to the discretization of the cloth image. Alternatively, wrinkles may be computed, and their local geometry provides valuable information not only about a convenient position for the gripper, but also about its optimal orientation. These methods are computationally more intensive (their complexity depending on the texture and actual wrinkles, local computations being more costly than for the segmentation-based methods) but their outcome is more reliable and informative.

Each one of the three variants of **specific GPL** has its own difficulties. Corner and edge detection methods have to distinguish these features from folds that may appear on the outline of the cloth image (some references also detect corners which appear inside the

cloth patch, as a result of a fold). The complexity is that of the polygonal approximation of the outline, as well as the number of folds and wrinkles that may be mistaken for real edges and corners. Some procedures for disambiguation have been described. A shortcut consist in resorting to the lowermost point, if the cloth is hanging. Specific garment features require a previous training step, and sophisticated classification algorithms. The complexity depends here on the specific features taken for the learning phase, i.e. the dimension of feature vectors (and the computational intricacy of its elements) and the number of samples needed for obtaining a reasonable classifier. Nodes from models, finally, quite related to classification and state estimation methods (if not identical at all), are difficult precisely for the required model matching step. The complexity is conditioned, in this case, by the discretization of the model (number of nodes or of triangles in a mesh model). In all of the three variants, failure in GPL leads either to no grasp or to grasp another feature than the intended one. However, specific GPL usually forms part of a repeated grasp and classification/state estimation strategy, which means that such failures can be corrected in successive regrasping steps.

In a somehow generic and abstract way, one could be tempted to attribute the difficulty of **classification** to inter-garment variability, and that of **state estimation** to intra-garment variability. Here, difficulty is associated to complexity: there are many types and variants of garments, and an infinity of possible states. However, as classification and state estimation are often performed simultaneously, things are not as straightforward: in shape recognition-based methods, classification is difficult because in some states one type of cloth can be mistaken for another (i.e., they are indistinguishable). In this same category of methods, state estimation faces the additional problem of discretizing the continuous space of deformations of cloth. Solutions to this problem, as described in the survey, include either resorting to easy-to-compute specific geometric measures of the outline (complexity is conditioned by the number of such features to compare between observations and models), or by associating a mesh to each garment type, where the nodes implicitly define discrete states when the robot grasps the cloth item at each one of them and lifts the cloth. The complexity of matching depends obviously on the granularity of the mesh. Either planar projections of such meshes are compared or directly 3D representations, by a variety of algorithms that range form simple overlap criterion to sophisticated dynamic programming methods, POMDPs strategies or CNNs, passing through nearest neighbor or SVM-based methods.

Alternatively to shape recognition-based approaches, classification and state estimation can be performed by aiming at specific garment features like collars, hemlines or cuffs. This has been little explored, and mainly for classification purposes. Although here it is not necessary to consider the whole shape of the cloth item, the search of the identifying features may also be costly (besides the need of a previous learning phase of such features). State estimation requires an additional step of determining the relative spatial relationships between localized features.

The following tables provide a summary on the references of the specific GPL methods and Classification and State estimation algorithms, attending to the structuring into subsections and stressing other factors like the presentation of the cloth items. In particular, Table 9 distributes the references discussed in Section 3.2 according to the type of feature identified by the vision system and the location of the cloth item.

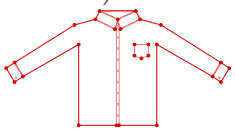
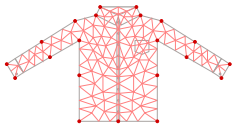
	Generic (edges, corners) 	Specific features 	Nodes on model 	Fiducial markers
Lying on surface	[22] [39] [40]	[43, 45]	[46]	
Hanging in the air	[29] [37, 38] [25] [26]	[41]	[2, 36, 48, 49, 30] [47]	[31]

Table 9: Sorting of references according to the feature type the vision system aims at and whether the cloth item appears lying on a surface or is hanging in the air, grasped by the robot.

Table 10 groups together the references explained in Section 4 according to the state of the cloth item and the classification data type used.

The reviewed contributions have been distributed according to the structure of this survey. Coherence within a subject-centered presentation has required to split the contents of some referenced works that deal with the different topics. For example, the graspability graph construction for grasp node identification of the reconfiguration phase in [23] is described in Section 3.2.3, whereas their contour matching algorithm for classification is explained in Section 4.1.2 and the corresponding active sensing strategy in Section 4.1.4. This allows to evaluate the different alternatives to each of the subproblems more easily than in a reference-guided structure or an annotated bibliography.

By observing Table 10 it becomes evident that the chosen identification method (i.e., shape-based or feature-based) depends on how the cloth item is presented to the vision system. If the garment is in a canonical configuration (e.g., totally extended on a surface or hanging while grasped at specific points like the shoulders of a shirt) then the overall shape is clearly most identifying for the type of clothing and its state. On the other hand, bunched or crumpled states make it more advisable to resort to local features, which generally are easier to recognize in such states than the overall shape. Such local features (collars, hemlines or openings, buttons, pockets, cuffs, heels, etc., we adopt the term *characteristics* of [51]) are more versatile, as for the presentation of the cloth item, with the only requirement for such identifying characteristics to be at least partially visible. However, such characteristics may be costly to detect, and their mere identification only solves the classification problem (if they are discriminating enough), whereas state recognition requires in addition to determine the mutual spatial relationship of such features within the cloth item. Up to date, this has only been solved for a simple category, socks, where the relative location of tip, heel and opening determines the state [52].

More basic and generic features like SIFT features do not have an intuitive semantic interpretation, but require less processing steps. Other basic visual clues like color can hardly be used for classification, as for robotic manipulation purposes: a shirt is a shirt independently of whether it is black, white or red, and it will be folded in the same way –or the procedure to assist a disabled person to put it on will be the same– in any case. However, it may be relevant in certain applications like matching of pairs of socks, or when for some reason the test set is bounded and the colors of the different garments are

Based on	Matching	Data	Cloth presentation		
			Hanging	Crumpled	Loosely extended
Shape	Geometric features	Synthetic	[24]		
		Real	[27]$^{(2)}$		
	Contour matching	Synthetic	[54, 2]$^{(1)}$ [23]		
		Real	[25] [27]$^{(2)}$	[52]	[4]$^{(2)}$[52][53, 5]
	3D matching	Synthetic	[36, 48, 49, 30]$^{(1)}$ [55] (SIFT) [56] (occupancy) [59](raw depth)		
		Real	[41, 57][59][60]	[58]	
Features	Generic	Real		[50]$^{(2)}$ (wrinkles)	
	Characteristics	Real		[52] [51]$^{(2)}$	[52]

Table 10: Overview on classification and state estimation references, according to the overall state in which the cloth item is presented to the vision system vs. the type of data used for classification. Some references appear more than once, as they resort to different methods. (1) Only state estimation, the garment type is assumed to be known beforehand. (2) Only classification, the state is assumed to be known or is irrelevant for the task.

known beforehand and are discriminative. Identifying the fabric type e.g. through the wrinkle appearance [50] may be also discriminative in limited sets, it is unlikely that a shirt be made of knitted fabrics or underwear made of denim, but it is not impossible. Furthermore, different garments may be made of the same fabrics. New descriptors, like the Deformation and Light Invariant (DaLI) [61] inspired on the Heat Kernel Signature (HKS), are quite suitable for clothing under varying photometric conditions, due to their resilience under non-rigid transformations, and could be further explored in the context of the visual processes in robot cloth manipulation.

Classification and state estimation are quite suitable for benchmarking. The creation of an annotated dataset of clothes to provide a unified base against which to compare the different methods is highly valuable. Such databases should include different garment types with collections of individuals in different states, and should reflect the alternatives of performing categorization both shape- or feature-based. One of the most complete and recent efforts in creating a garment database for robot manipulation is presented in [45], who include appearance and depth data, and annotate garment parts like sleeves, collars, hips or hemlines. This dataset has been developed in the context of informed grasp point localization, but it could be used for classification as well (in fact, garment parts are

associated to specific garment types). The authors review the few other garment datasets existing in the literature, mostly aimed at classification, but also oriented to folding, cloth segmentation, flattening, and the like [62, 63, 64, 51, 50, 41].

Besides better efficiency and robustness, the desideratum for future work in classification and state recognition, from our point of view, should aim at a higher degree of versatility as for possible categories. Such new states could include partial or total reversals of the cloth part (i.e., a shirt's sleeve inside-out) or, for the specific application of clothing assistance, the different phases of clothing (or unclothing) a person, to assess the degree of completion of the task.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under projects RobInstruct (TIN2014-58178-R) and I-DRESS: Assistive interactive robotic system for support in dressing (PCIN-2015-147). I would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the survey.

[1] Nadia Magnenat-Thalmann and Pascal Volino. From early draping to haute couture models: 20 years of research. *The Visual Computer*, 21(8-10):506–519, 2005.

[2] Yasuyo Kita, Fuminori Saito, and Nobuyuki Kita. A deformable model driven visual method for handling clothes. In *2004 IEEE International Conference on Robotics and Automation, ICRA 2004, New Orleans (LA), USA, April, 2004*, pages 3889–3895, 2004.

[3] Saúl Cuen, Juan Andrade-Cetto, and Carme Torras. Action selection for robotic manipulation of deformable objects. In *Proc. of the ESF-JSPS Conf. on Experimental Cognitive Robotics*, march 2008.

[4] Stephen Miller, Mario Fritz, Trevor Darrell, and Pieter Abbeel. Parametrized shape models for clothing. In *International Conference on Robotics and Automation (ICRA)*, pages 4861–4868, 2011.

[5] Jan Stria, Daniel Průša, Vaclav Hlaváč, Libor Wagner, Vladimír Petrík, Pavel Krsek, and Vladimír Smutný. Garment perception and its folding using a dual-arm robot. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 61–67, Sept 2014.

[6] Gualtiero Fantoni, Marco Santochia, Gino Dinia, Kirsten Tracht, Bernd Scholz-Reiter, Juergen Fleischer, Terje Kristoffer Lien, Guenther Seliger, Gunther Reinhart, Joerg Franke, Hans Nrgaard Hansen, and Alexander Verl. Grasping devices and methods in automated production processes. *CIRP Annals - Manufacturing Technology*, 2014.

[7] David Gershon. Parallel process decomposition of a dynamic manipulation task: robotic sewing. *Robotics and Automation, IEEE Transactions on*, 6(3):357–367, Jun 1990.

[8] Panagiotis N. Koustoumpardis, Paraskevi Zacharia, and Nikos A. Aspragathos. Intelligent robotic handling of fabrics towards sewing. In *Industrial Robotics*, chapter 28, pages 559–581. 2006.

[9] Paraskevi Th. Zacharia, Nikos A. Aspragathos, Ioannis G. Mariolis, and Evangelos Dermatas. A robotic system based on fuzzy visual servoing for handling flexible sheets lying on a table. *Industrial Robot*, 36(5):489–496, 2009.

[10] Paraskevi Zacharia. Robot handling fabrics towards sewing using computational intelligence methods. In Dr. Ashish Dutta, editor, *Robotic Systems - Applications, Control and Programming*, pages 61–84. 2012.

[11] Georgios Zoumponos and Nikos Aspragathos. A fuzzy strategy for the robotic folding of fabrics with machine vision feedback. *Industrial Robot*, 37(3):302–308, 2010.

[12] Kimitoshi Yamazaki, Ryosuke Oya, Kotaro Nagahama, and Masayuki Inaba. A method of state recognition of dressing clothes based on dynamic state matching. In *System Integration (SII), 2013 IEEE/SICE International Symposium on*, pages 406–411, Dec 2013.

[13] Tomoya Tamei, Takamitsu Matsubara, Akshara Rai, and Tomohiro Shibata. Reinforcement learning of clothing assistance with a dual-arm robot. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 733–738, Oct 2011.

[14] Nishanth Koganti, Tomoya Tamei, Takamitsu Matsubara, and Tomohiro Shibata. Estimation of human cloth topological relationship using depth sensor for robotic clothing assistance. In *Proceedings of Conference on Advances In Robotics*, AIR '13, pages 36:1–36:6, New York, NY, USA, 2013. ACM.

[15] Kimitoshi Yamazaki and Masayuki Inaba. A cloth detection method based on image wrinkle feature for daily assistive robots. In *Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA 2009), Keio University, Yokohama, Japan, May 20-22, 2009*, pages 366–369, 2009.

[16] Kimitoshi Yamazaki, Ryohei Ueda, Shunichi Nozawa, Yuto Mori, Toshiaki Maki, Naotaka Hatao, Kei Okada, and Masayuki Inaba. System integration of a daily assistive robot and its application to tidying and cleaning rooms. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, pages 1365–1371, 2010.

[17] Kimitoshi Yamazaki, Kotaro Nagahama, and Masayuki Inaba. Daily clothes observation from visible surfaces based on wrinkle and cloth-overlap detection. In *Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA 2011), Nara Centennial Hall, Nara, Japan, June 13-15, 2011*, pages 275–278, 2011.

[18] K. Paraschidis, Nikolaos Fahantidis, Vassilios Petridis, Zoe Doulgeri, Loukas Petrou, and Georgios Hasapis. A robotic system for handling textile and non rigid flat materials. *Computers in Industry*, 26(3):303–313, August 1995.

[19] Li Sun, Gerardo Aragon-Camarasa, Simon Rogers, and J.Paul Siebert. Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 185–192, May 2015.

[20] Jan J. Koenderink and Andrea J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557 – 564, 1992.

[21] Li Sun, Gerardo Aragon Camarasa, Aamir Khan, Simon Rogers, and Paul Siebert. A precise method for cloth configuration parsing applied to single-arm flattening. *International Journal of Advanced Robotic Systems*, 13, April 2016. CLOPEMA - 288553 (Clothes Perception and Manipulation).

[22] Eiichi Ono, Noboyuki Sakane, and Shigeyuki Sakane. Unfolding folded fabric using outline information with vision and touch sensors. *Journal of Robotics and Mechatronics*, 10(3):235–243, 1998.

[23] Marco Cusumano-Towner, Arjun Singh, Stephen Miller, James F. O'Brien, and Pieter Abbeel. Bringing clothing into desired configurations with limited perception. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA) 2011*, pages 1–8, May 2011.

[24] Manabu Kaneko and Masayoshi Kakikura. Planning strategy for putting away laundry -isolating and unfolding task. In *Proc. of the 4th IEEE Int. Symposium on Assembly and Task Planning*, pages 429–434, 2001.

[25] Fumiaki Osawa, Hiroaki Seki, and Yoshitsugu Kamiya. Unfolding of massive laundry and classification types by dual manipulator. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, pages 457–463, 2007.

[26] Jeremy Maitin-Shepard, Marco Cusumano-Towner, Jinna Lei, and Pieter Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *International Conference on Robotics and Automation (ICRA*, pages 2308–2315, 2010.

[27] Brian Willimon, Stan Birchfield, and Ian Walker. Classification of clothing using interactive perception. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1862–1868, May 2011.

[28] Kyoko Hamajima and Masayoshi Kakikura. Planning strategy for unfolding task of clothes – isolation of clothes from washed mass. In *SICE'96*, volume 2, pages 1237–1242, 1996.

[29] Kyoko Hamajima and Masayoshi Kakikura. Planning strategy for task of unfolding clothes. *Robotics and Autonomous Systems*, 32(2–3):145–152, 2000.

[30] Yasuyo Kita, Fumio Kanehiro, Toshio Ueshiba, and Nobuyuki Kita. Clothes handling based on recognition by strategic observation. In *11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011), Bled, Slovenia, October 26-28, 2011*, pages 53–58, 2011.

[31] Christian Bersch, Benjamin Pitzer, and Sören Kammel. Bimanual robotic cloth manipulation for laundry folding. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1413–1419, Sept 2011.

[32] Seiji Hata, Takehisa Hiroyasu, Junichiro Hayashi, Hirotaka Hojoh, and Toshihiro Hamada. Robot system for cloth handling. In *34th Annual Conference of the IEEE Industrial Electronics Society, IECON*, pages 3449–3454, 2008.

[33] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. Determining where to grasp cloth using depth information. In *Artificial Intelligence Research and Development - Proceedings of the 14th International Conference of the Catalan Association for Artificial Intelligence, Lleida, Catalonia, Spain, October 26-28, 2011*, pages 199–207, 2011.

[34] Guillem Alenyà, Arnau Ramisa, Francesc Moreno-Noguer, and Carme Torras. Characterization of textile grasping experiments. In *Proceedings of the 2012 ICRA Workshop on Conditions for Replicable Experiments and Performance Comparison in Robotics Research*, pages 1–6, St Paul (MN), 2012.

[35] Pol Monso, Guillem Alenyà, and Carme Torras. Pomdp approach to robotized clothes separation. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pages 1324–1329, 2012.

[36] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. Clothes state recognition using 3d observed data. In *2009 IEEE International Conference on Robotics and Automation, ICRA 2009, Kobe, Japan, May 12-17, 2009*, pages 1220–1225, 2009.

[37] Khairul Salleh, Hiroaki Seki, and Yoshitsugu Kamiya. Tracing manipulation in clothes spreading by robot arms. *J Rob Mechatron*, 18(5):564–571, 2006.

[38] Khairul Salleh, Hiroaki Seki, Yoshitsugu Kamiya, and Masatoshi Hikizu. Inchworm robot grippers for clothes manipulation. *Artificial Life and Robotics*, 12(1–2):142–147, march 2008.

[39] Brian. Willimon, Stan Birchfield, and Ian Walker. Model for unfolding laundry using interactive perception. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 4871–4876, Sept 2011.

[40] Dimitra Triantafyllou and Nikos A. Aspragathos. A vision system for the unfolding of highly non-rigid objects on a table by one manipulator. In *Intelligent Robotics and Applications - 4th International Conference, ICIRA 2011, Aachen, Germany, December 6-8, 2011, Proceedings, Part I*, pages 509–519, 2011.

[41] Andreas Doumanoglou, Andreas Kargakos, Tae-Kyun Kim, and Sotiris Malassiotis. Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 987–993, May 2014.

[42] Dimitra Triantafyllou, Ioannis Mariolis, Andreas Kargakos, Sotiris Malassiotis, and Nikos Aspragathos. A geometric approach to robotic unfolding of garments. *Robot. Auton. Syst.*, 75(PB):233–243, January 2016.

[43] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. Using depth and appearance features for informed robot grasping of highly wrinkled clothes. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1703–1708, May 2012.

[44] Peter Gibbons, Phil Culverhouse, and Guido Bugmann. Visual identification of grasp locations on clothing for a personal robot. In *Proceedings of Taros'09*, pages 78–81, Londonderry, 2009.

[45] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. Learning rgb-d descriptors of garment parts for informed robot grasping. *Engineering Applications of Artificial Intelligence*, 35:246 – 258, 2014.

[46] Stephen Miller, Jur van den Berg, Mario Fritz, Trevor Darrell, Kenneth Y. Goldberg, and Pieter Abbeel. A geometric approach to robotic laundry folding. *International Journal of Robotic Research*, 31(2):249–267, 2012.

[47] Yinxiao Li, Danfei Xu, Yonghao Yue, Yan Wang, Shih-Fu Chang, Eitan Grinspun, and Peter K. Allen. Regrasping and unfolding of garments using predictive thin shell modeling. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 1382–1388, 2015.

[48] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. A method for handling a specific part of clothing by dual arms. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 11-15, 2009, St. Louis, MO, USA*, pages 4180–4185, 2009.

[49] Yasuyo Kita, Ee Sian Neo, Toshio Ueshiba, and Nobuyuki Kita. Clothes handling using visual recognition in cooperation with actions. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, pages 2710–2715, 2010.

[50] Kimitoshi Yamazaki and Masayuki Inaba. Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, November 3-7, 2013*, pages 2710–2717, 2013.

[51] Bryan Willimon, Ian Walker, and Stan Birchfield. Classification of clothing using midlevel layers. *ISRN Robotics*, 2013.

[52] Ping Chuan Wang, Stephen Miller, Mario Fritz, Trevor Darrell, and Pieter Abbeel. Perception for the manipulation of socks. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2011, San Francisco, CA, USA, September 25-30, 2011*, pages 4877–4884, 2011.

[53] Jan Stria, Daniel Průša, and Václav Hlaváč. Polygonal models for clothing. In Michael Mistry, Aleš Leonardis, Mark Witkowski, and Chris Melhuish, editors, *Advances in Autonomous Robotics Systems*, volume 8717 of *Lecture Notes in Computer Science*, pages 173–184. Springer International Publishing, 2014.

[54] Yasuyo Kita and Nobuyuki Kita. A model-driven method of estimating the state of clothes for manipulating it. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, WACV '02, pages 63–69, Washington, DC, USA, 2002. IEEE Computer Society.

[55] Yinxiao Li, Chih-Fan Chen, and Peter K. Allen. Recognition of deformable object category and pose. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 5558–5564, May 2014.

[56] Yinxiao Li, Yan Wang, M. Case, Shih-Fu Chang, and Peter K. Allen. Real-time pose estimation of deformable objects using a volumetric approach. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 1046–1052, Sept 2014.

[57] Andreas Doumanoglou, Tae-Kyun Kim, Xiaowei Zhao, and Sotiris Malassiotis. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, chapter Active Random Forests: An Application to Autonomous Unfolding of Clothes, pages 644–658. Springer International Publishing, Cham, 2014.

[58] Li Sun, Simon Rogers, Gerardo Aragon-Camarasa, and J.Paul Siebert. Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2464–2470, May 2016.

[59] Ioannis Mariolis, Georgia Peleka, Andreas Kargakos, and Sotiris Malassiotis. Pose and category recognition of highly deformable objects using deep learning. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 655–662, July 2015.

[60] Christos Kampouris, Ioannis Mariolis, Georgia Peleka, Evangelos Skartados, Andreas Kargakos, Dimitra Triantafyllou, and Sotiris Malassiotis. Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1656–1663, May 2016.

[61] Edgar Simo-Serra, Carme Torras, and Francesc Moreno-Noguer. Dali: Deformation and light invariant descriptor. *International Journal of Computer Vision*, 115(2):136–154, 2015.

[62] Gerardo Aragon-Camarasa, Susanne Oehler, Yuan Liu, Sun Li, W. Paul Cockshott, and J. Paul Siebert. Glasgow's stereo image database of garments. *CoRR*, abs/1311.7295, 2013.

[63] Ioannis Mariolis and Sotiris Malassiotis. *Computer Analysis of Images and Patterns: 15th International Conference, CAIP 2013, York, UK, August 27-29, 2013, Proceedings, Part II*, chapter Matching Folded Garments to Unfolded Templates Using Robust Shape Analysis Techniques, pages 193–200. Springer, Berlin, Heidelberg, 2013.

[64] Libor Wagner, Daniela Krejčová, and Vladimír Smutný. CTU color and depth image dataset of spread garments. Research Report CTU–CMP–2013–25, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, September 2013.