Clothes Perception and Manipulation

# D5.1

# Report on the Reasoning Mechanism and the Object Representation

Version: 1.0

Last Update: 31-1-2013

Distribution Level: PU (Public)

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement *n*$^{o}$ 288553

| | | |
|---|---|---|
| **Project Number** | : | FP7-288553 |
| **Project Title** | : | Clothes Perception and Manipulation |

| | | |
|---|---|---|
| **Deliverable Number** | : | D5.1 |
| **Title of Deliverable** | : | Report on the Reasoning Mechanism and the Object Representation |
| **Nature of Deliverable** | : | Report |
| **Dissemination Level** | : | Public |
| **Version** | : | 1.0 |
| **Contractual Delivery Date** | : | 31-1-2013 |
| **Actual Delivery Date** | : | 31-1-2013 |
| **Contributing WP** | : | WP5 |
| **Author(s)** | : | Dimitra Triantafilou, Andreas Doumanoglou, Ioannis Mariolis, Dimitris Gorpas, Sotiris Malassiotis (CERTH), Libor Spacek (CMP) |
| **Reviewer(s)** | : | Libor Spacek, Sotiris Malassiotis |

## Abstract

This report describes techniques for robust extraction of persistent features such as folds, corners and edges from a noisy range image, followed by connecting these features into a topological structure that is mapped to a garment template by using knowledge and geometry. The resulting mapping allows detection of grasping affordances or planning robot movements that will reveal such affordances.

Algorithms for garment shape recognition in a flat state are presented. The algorithms are designed to be robust to moderate deformations.

Another algorithmic method recognizes a garment and registers its shape with a template even when the garment is folded on itself. The folding axis is also identified with respect to the template, thus allowing the planning of an unfolding sequence.

First experiments exploiting only boundary shape primitives demonstrate promising results under significant occlusions of garment parts.

Finally, a linguistic inspired architecture based on graph grammars is proposed to perform both bottom-up and top-down parsing of a garment image incorporating high-level knoweldge of the garment structure.

**Keywords**

garment manipulation, visual scene understanding, perception

# Revision History

| Version | Description | Author(s) | Date |
|---|---|---|---|
| 0.1 | First Draft | CERTH | 7 Jan 2013 |
| 0.8 | Early Final Draft | CERTH | 18 Jan 2013 |
| 0.9 | Corrected Version | Libor Spacek,CMP | 31 Jan 2013 |
| 1.0 | Final Version+Abstract | Libor Spacek,CMP | 1 Feb 2013 |

# Contents

# 1 Introduction

Clothes manipulation requires advanced visual and dexterity skills. Children manage to perform clothes manipulation tasks such as folding a towel only when they reach an age of between 3 and 4 years old. Due to the soft and flexible nature of the manipulated materials, the visual system has to work hard dealing all the time only with sparse visual cues. This is because the cloth surface may be mostly occluded and the 3D shape of the garment object may be deformed to a configuration hardly resembling its prototypical shape.

Yet, by watching humans performing tasks such as laundry folding, one may learn a more or less consistent strategy of manipulation steps that have the goal of bringing the item into a progressively simpler configuration:

- *isolating* an item from the heap by picking and lifting
- *flattening*, i.e. iterating a series of picking and lifting actions by alternating hands that will not allow the garment to fall. Alternatively, one may slide the fingers along a seam of the garment while holding it at another point on the seam. In this step the goal is the identification of two grasp points that will bring the garment into a relatively "flat" hanging configuration (under gravity). In this state the garment will exhibit simple folds, largely resembling layers of planar polygons. After this step the garment may be spread out and laid on a flat surface such as a table to free both hands
- complete *unfolding* of the garment, using one or both hands
- a sequence of stereotypical movements for*folding* the garment.

Prior learning is also essential in performing this task accurately and without resorting to blind exploration. Recognition of the type of the garment (e.g. towel, shirt, trousers) will guide, for example, the search for the grasp points during the flattening stage. Towels should be ideally held by two non-opposing corners, and shirts should be ideally held near the top of the sleeves. Unfolding the garment is effective only when the goal shape is known, and folding it requires that structural features such as buttonline, neckline and the front side have been identified. Apart from recognition, elementary physics will be learned from experience. For example, we may roughly predict the shape of a towel, held by two non-opposing corners, colliding with a table. This allows placing the towel on the table without any folds.

What will it take for a robot to perform the laundry task? Only a few researchers have explored this question during the last couple of years. Although successful demonstrations have already been achieved by US and Japanese groups, they are characterized by: a) strong simplifying assumptions to facilitate visual perception and b) exploiting only low-level visual cues and relatively "blind" manipulation (e.g. rotating a hanging towel until a corner is found). These restrictions hinder the applicability of these techniques to the real-world conditions.

From the point of view of required computer vision skills, *CloPeMa* will work towards: a) advancing the state-of-the-art in visual perception of deformable materials such as clothes and b) incorporation of high-level knowledge and reasoning in different levels of the perception chain. This deliverable presents the first achievement towards these goals.

In section 2 we briefly review the state-of-the-art in visual perception and modeling of deformable objects and try to link it with clothes perception.

In section 3 we try to explore the question whether the configuration of a cloth hanging from a single point may be recovered from a single image. Therefore we developed techniques for robust extraction of persistent features such as folds, corners and edges from a noisy range image. Then we connect these features into a topological structure that we are able to map to the cloth rectangular template by imposing prior-level knowledge and geometry reference. The resulting mapping allows us to detect grasping affordances or if none is identified to intelligently plan a movement that will reveal such affordances.

In section 4 we present algorithms for garment shape recognition in a flat state. The algorithms are designed to be robust to moderate deformations.

We also present an algorithm (section 5) that will recognize a garment and will register its shape with a template even if the garment is folded on itself. The folding axis is also identified with respect to the template thus allowing for planning an unfolding sequence.

Finally, in section 6 we propose a linguistic inspired architecture based on graph grammars that is able to perform bottom-up, top-down parsing of a garment image incorporating high-level knowledge of the garment structure.

Very first experiments exploiting only boundary shape primitives demonstrate promising results even under significant occlusions of garment parts.

# 2 State of the Art

We briefly review the state-of-the-art in computer vision research related to modelling, reconstruction and recognition of deformable objects. A state-of-the-art of vision tasks related to clothes manipulation will be given in the individual sections of proposed algorithms.

## 2.1 Physics Based Modelling

Physics based modelling techniques use the laws of physics to simulate the shape and movement of deformable objects such as clothes. Terzopoulos et al. [55] in the 90s introduced the concept that has since become widespread in the computer graphics community. Later on several applications in computer vision and especially medical image processing used the concept of estimating surfaces deforming under physics laws. Physics-based models of garments are today applied widely for realistically simulating virtual humans in various applications and cloth simulation is today integrated into both commercial and open source 3D modelling applications (e.g. Blender). Industrial applications of cloth and garment simulation techniques include material simulation and product design, film and 3D game production and medical simulation.

Physically-based cloth simulation is commonly formulated as a partial differential equation over time:

$$\ddot{x} = M^{-1}(-\frac{\partial E}{\partial x} + F) \tag{1}$$

where $x$ is the shape of the cloth and $M$ is its mass distribution. $E$ is the cloth's internal energy such as stretching, bending and shearing and $F$ are the external forces (as a function of $x$ and $\dot{x}$ ) which include rigid attachment points, collision and self-collision forces, damping etc.

This is subsequently discretized using a mass-spring model or finite-element model and the resulting differential equations are solved using numerical integration techniques. Early methods used explicit integration techniques (such as Euler) [8] which however require many small time steps to stably advance the simulation. More recent implicit [2] and semi-implicit [16] integration techniques on the other hand are more suitable for stretch-resistant cloth physics.

Current research on physics-based modelling of cloths is aiming at the following research challenges [11]: a) increasing the accuarcy and realism of the simulation by handling non-linearity and exploiting more accurate finite-element models, b) increasing the speed of the simulation without sacrificing the accuracy or resolution, c) development of efficient collision and self-collision detection as well as collision response generation techniques, d) linking the parameters of mathematical models with real-world material properties, thus allowing for easier control of the simulation fidelity.

In the field of computer vision, the term physics-based modelling is used to describe techniques that exploit physical laws for understanding the appearance and movement of objects in images. Most well known such techniques are “Snakes” [29], simulated elastic strings, that are widely used for image segmentation. Many such techniques have also been developed for the analysis of image sequences of deformable objects [26, 38]. Most of these works are suitable for objects depicted in relatively simple deformations, such as bending.

On the other hand, little effort has been put into employing physics-based models to track cloth motion. The majority of few existing approaches employ an analysis-by-sythesis approach seeking to minimize with respect to unknown parameters the discrepancy between the actual image of the fabric and one rendered based on cloth simulation techniques.

Jojic and Huang [27] register 3D range data of a known rectangular piece of cloth laid over a known collision object with the corresponding simulated geometry in order to obtain the simulation model’s parameters that lead to the depicted real-world draping. Similarly, Bhat et. al [6] used image sequences of real fabrics under motion to calibrate the static and dynamic parameters of a cloth simulator. They compare two video sequences, one from simulation and one from the real world, by means of the discrepancy in the detected folds as well as the depicted silhouettes. Charfi et al.[9] estimate the damping parameters of a textile by analysing a drooping cloth. Their approach is based on analyzing the trajectory of markers on the surface of the cloth and minimizing their discrepancy with respect to those produced by a cloth simulator. A markerless approach is used in [23] for motion tracking and parameter estimation of a known fabric.

The above physics-based cloth estimation techniques do not address the problem of estimation of the configuration of a possibly unknown garment, constrained by one or more points. For known cloth geometry Kitta et. al [32] use a database of previously simulated cloth configurations (of a known item) hanged from different points. This is used to determine the state (grasping point) of the cloth depicted in an image by searching the database for an image with a similar silhouette. A success rate of 80% is reported. Similarly, Cusumano et. al [15] use a cloth simulator to obtain a generative model of the item’s sihlouette given the hanging points and the item’s prototypical shape (from a small set).

## 2.2 Deformable Templates

A simple but intuitive representation of cloth geometry is a quasi-developable surface or a surface isometrically embedded in 3D. More simply a planar polygon surface or 2D template shape may be deformed to 3D space using a quasi-isometric transformation. We call such a representation a deformable template. The key problem here is the recovery or reconstruction of the surface given an image of the cloth under random configuration. The problem is commonly addressed as a problem of finding dense correspondences between image points and template points. These are subsequently used to construct a parametrization of the discrete surface so that all points may be mapped to points on the template. In the context of robotic cloth manipulation, this approach would be very useful for generating action plans that will lead the item to an unfolded configuration.

Should the input contain a 2D image or texture image of the cloth or garment, the first step in surface recovery, correspondence estimation, is usually simplified by assuming that the surface is covered by a non-repetitive texture. Then local feature descriptors such as SIFT may be used for robust correspondence finding [45, 13, 1]. Correspondence filtering is achieved by imposing isometric constraints among candidate correspondences as in [45, 13]. Both approaches above exploit local consistency in candidate matches. In [13] clusters in matched features may be used to segment the surface into developable patches using a region growing strategy. In [58] a checkerboard pattern printed on a cloth is tracked from a video sequence thus allowing for easier correspondence estimation and surface parametrization. Similarly, [52] use a specially designed color coded pattern. Local techniques [45, 13] or global techniques [19, 51] are subsequently used to obtain the surface parametrization.

Several works also use the assumption that the deformable surface is planar on the local scale [54, 12, 50]. Surface patches are reconstructed first and glued together in a second step. Others [7, 3] will explicitly model the underlying template warping function and reconstruct its control parameters by imposing the isometry constraint. The warping function is assumed to be continuously differentiable and invertible. In contrast [20] propose analytical solutions of the warping function given the boundary contour only.

There are several limitations of the state-of-the-art in the context of robotic manipulation:

- although the global shape of the template may be assumed to be known, its dimensions and scale are usually unknown. So is the texture of the garment, if any. This makes techniques relying on dense point correspondences difficult to apply

- existing techniques will be more suitable for limited deformations such as a waving flag or a bent piece of paper. Most of the existing works do not deal with occlusions caused by folds and when they do they consider them as outliers. Thus the occlusions coverage is assumed to be small

- existing techniques assume most of the surface is visible, which is not the case with a piece of cloth hanging from one or two points. Partial matching techniques would be useful in this situation.

## 2.3 Part-based Representations

In the previous section we reviewed techniques that try to capture a global deformation function from images. An alternative approach is to use independent models for local appearance over parts or patches and a geometric model for the spatial relationship among the parts. This approach was proven quite promising especially in situation where local appearance may be modelled effectively (e.g. faces or body parts) and geometric variability may be learned from examples.

Existing techniques mainly differ in the way they impose geometric constraints and the techniques used to fit the model to the image. Constellation model techniques constrain the part locations into a fixed set of configurations (e.g. using interest point detectors) and subsequently learn Gaussian distributions of part relationships (e.g. [18]) resulting in a generative model for each object class. Since each detected feature may be assigned to one of possible parts, all permutations of features over parts need to be examined to determine object class and configuration, thus being computationally expensive. The Pictorial Structure approach e.g. [17] explicitly models part relationship by a graph structure with known topology. In this case a match cost is defined over a dense set of locations, and the geometric arrangement is captured by "springs" connecting the parts. By restricting the topology of the graph to a tree (star-like or fan-like) efficient algorithms based on dynamic programming and generalized distance transform may be applied for efficiently minimizing the energy of the model given the image.

While the above part-based deformable models can capture significant variations in appearance, they are not suitable for modelling object categories with rich structure, containing for example a hierarchy of parts, optional parts, or parts that themselves may belong to different categories. Grammar based models may be considered as generalization of deformable part models by representing objects using variable hierarchical structures. Each part in a grammar based model can be defined directly or in terms of other parts using a sequence of high-level production rules, thus allowing for explicitly modelling complex structural variations.

In the category of part-based representation we may also mention sparse or "bag-of-word" models. The majority of these techniques do not model the relationship between parts, nevertheless they may be quite effective especially when appearance cues over local patches provide strong discriminatory information.

# 3 Feature extraction and topological analysis of a hanging towel

## 3.1 Introduction

An important part of the robotic unfolding of garments is the determination of grasping points. The points from which the clothing article is grasped play decisive role in the formation of its configuration. Thus, their proper choice can accelerate and facilitate the unfolding procedure. On the other hand, the knowledge of the garment's current configuration provides useful information for the detection of the proper grasping points.

Three different approaches to the detection of grasping points for the unfolding task can be distinguished in the robotic literature. The first category detects characteristic features of the clothing article and uses them as grasping points. For example, in the case of a towel, its corners are considered good grasping candidates [15], [56], whereas for a polo-shirt it is its collar [46].

In the second approach the initial grasping points bring the garment in a more or less known configuration, so as to facilitate the further selection of grasping points. Osawa [44] and Cosumano-Towner [15] used a heuristic technique to achieve it. By iteratively grasping the lowest hanging points of the article, each kind of garment is led into a limited number of configurations (e.g. a towel would be grasped only by two non neighbouring corners). Similarly, Kaneko [28] detected parts of the hemline and re-grasped them so as the robot could hold the garment from two different edges of the hemline. In both cases, in the end, the garment is either unfolded or folded in half. The knowledge of the configuration facilitates the detection of the necessary grasping points for further unfolding manipulations, when needed.

Finally, other researchers estimate the garment's state, in order to detect the grasping points. Kita et al. [32] proposed a mass-spring model to simulate how clothing, in particular a T-shirt, hangs. The T-shirt is a priori grasped from a point on its hemline. Their work showed the ability of the simulator's models to adjust to the silhouettes of the true hanging garment while 3D point clouds extract the configuration of the clothing article with a good success rate. In addition, their system is able to identify and grasp a desired point with the other gripper. Bersch [5] used a T-shirt equipped with fiducial markers and transformed it from a random configuration into an unfolded state. At the end of the unfolding procedure the robot holds the T-shirt by its shoulders. The configuration of the T-shirt is estimated with the help of cloud representation and fiducial markers, while the selection of the next grasping point is based on a greedy policy that chooses the point with the smallest geodesic distance to the target grasping points.

In the presented approach, the topology of the visible features of a hanging towel is analysed. Firstly, the towel, which is hanging in the air from one of its corners, is decomposed into layers. Then, for each layer the corresponding position in an unfolded configuration of the towel is obtained. Our goal is to infer the possible configurations of the towel given only its visible part. Based on the set of obtained configurations we may facilitate the task of selecting grasp points or proposing a re-grasping strategy so that the grasp points will become visible.

## 3.2 Feature extraction

The first step for the topological analysis of the towel is the extraction of critical features such as edges and junctions that will provide all the information for the decomposition of the towel into layers. The extraction of these features and the way the towel is decomposed into layers is described in the two following subsections.

### 3.2.1 Detection of the towel's edges and junctions

The detection of the towel's edges and junctions is made in two phases. Firstly, the edges or parts of them are detected while in the second phase these edges are connected to each other in order to form longer edges or junctions. These features are extracted from 3D information of the towel's visible side, which is acquired from an Asus Xtion sensor.

So,initially, edge detection is performed on the 3D information of the towel. With the help of hough transform, the edges are separated into four categories: 1)the vertical, 2) the ones that incline rightwards, 3) the ones that incline leftwards and 4) the completely horizontal. Due to the noise of the data that are used and the curvature of the edges to be detected there is a tolerance on the inclination of the edges of each category. For example, in figure 1, all the vertical edges are marked with red. Next, all the pixels that are located close to each other and belong to edges that are classified into the same category are connected in order to form bigger edges (Douglas-Peucker is used in order to create polygonal lines from the pixels). In figure 2, the edges that are created after this procedure are depicted with red,green and blue for the categories 1,2 and 3 respectively.

At this phase, several edges of the towel are detected but it is not yet known how they are connected to each other. According to their classification, the edges might be connected to a bigger one or create a junction. Specifically, two edges of the same category can be unified, when the higher end of the first one is close to the lower end of the second one, or they can create a junction, when both lower/higher ends are close to each other. In

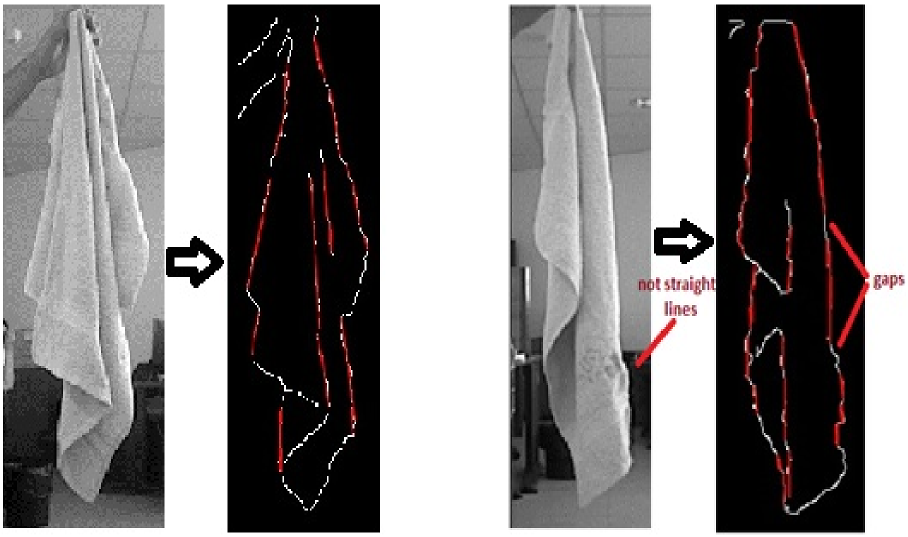


Figure 1: Detection of vertical edges

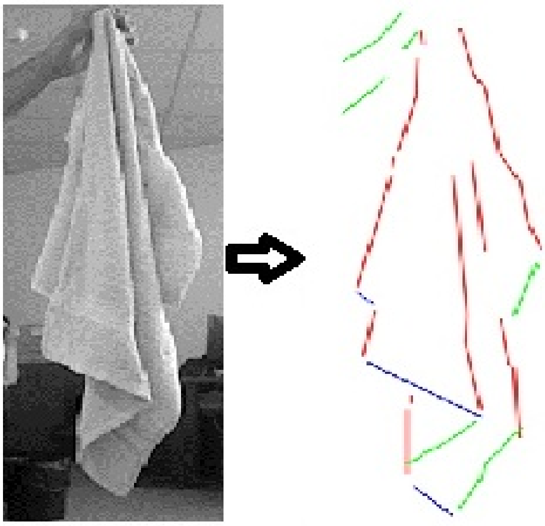


Figure 2: Detected edges of all the categories

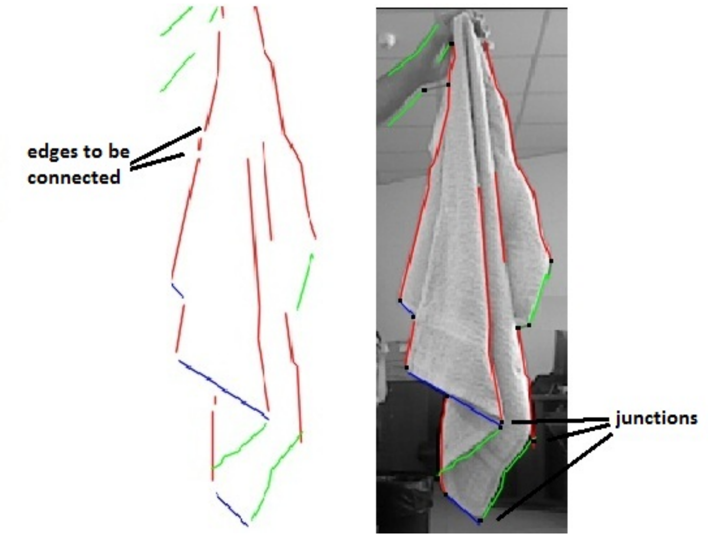


Figure 3: Junctions and connected edges

addition, when the end of an edge is close to another edge of a different category, then a junction is created. The edges' connection and the detection of junctions are depicted in figure 3.

### 3.2.2 Decomposition of the towel into layers

For the extraction of the towel's layers, the towel is held by one of its corners [1]. Its edges and junctions of edges are considered known. These features are either formed by folds or belong to the outline of the towel when it is fully extended. Folds, which are responsible for the formation of more than one layer, usually start close to the grasping point. In our method, all the edges that begin from a fold close to the grasping point are considered to start from it for convenience, without loss of generality.

The edges can be divided in two categories: 1) the vertical edges and, 2) the horizontal edges, which actually correspond to the categories 2,3 and 4 mentioned in the previous subsection (Fig 4.a). The vertical edges usually end at the grasping point. Sometimes, due to overlapping, there are vertical edges with the part close to the grasping point hidden from the camera. Nevertheless, for the formation of layers, the visible part, which is inclined towards the grasping point, is extended to it (Fig 4.b). In addition, there are vertical edges that might intersect far from the grasping point. These cases occur when a part of the towel is hidden behind the visible side(Fig 5.g).

The junctions that are taken into account for the extraction of layers are points of

[1] Our original intention was to start from random grasp points, but this was more challenging than expected and will be revisited in future

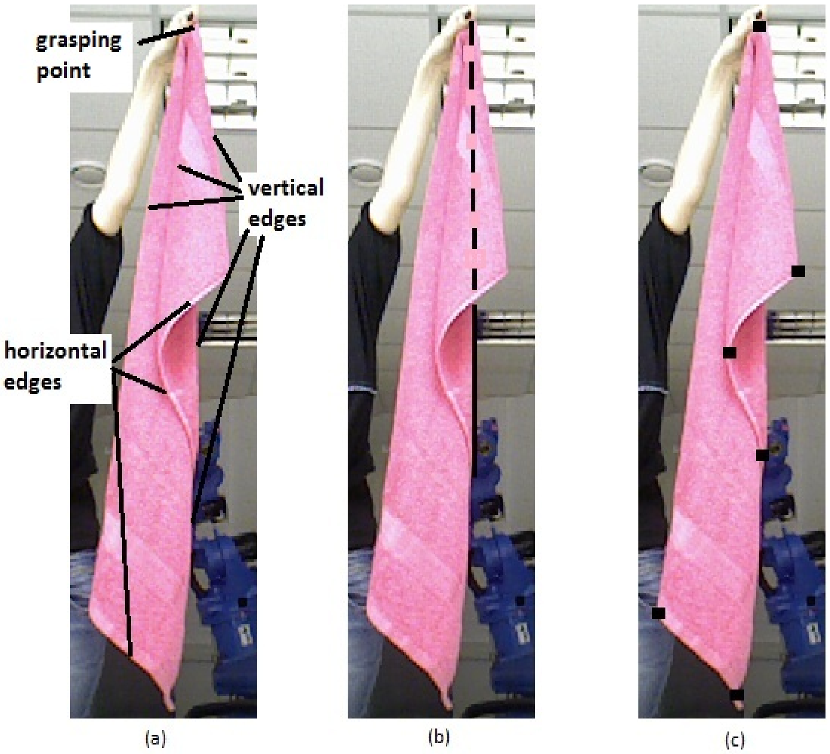


Figure 4: Edges and junctions of a hanging towel

intersection of two vertical or two horizontal lines or points where the end of a horizontal edge intersects with a vertical edge. In figure 4.c the junctions that are taken into account for the extraction of the towel's layers are depicted with small black squares.

For the extraction of the towel's layers, the following procedure is applied iteratively. The lowest corner, not examined yet, is picked and the two junctions it is connected to, that lead to higher located junctions, are examined. These junctions are connected to each other (in cases where the layer is triangular) or are both connected to the same junction (the layer is quadrangle). For example in figure 5.a, junction j1 is connected, through edges e2 and e3, to junctions j2 and j3, which are connected to each other with edge e3. So, a triangular layer, formed by the edges e1, e3 and e3, is detected. Similarly, a rectangular layer, formed by the edges e1, e2, e3 and e4, is detected in figure 5.d. If a junction is used as the lower junction of a layer then it cannot be used again in another layer. Actually, every junction can participate in N-1 layers, where N is the number of edges intersecting at this junction. The layers are represented by triangles or rectangles for convenience and in order to reduce computational complexity. In figure 5, the detected layers of two different configurations of the towel are highlighted.

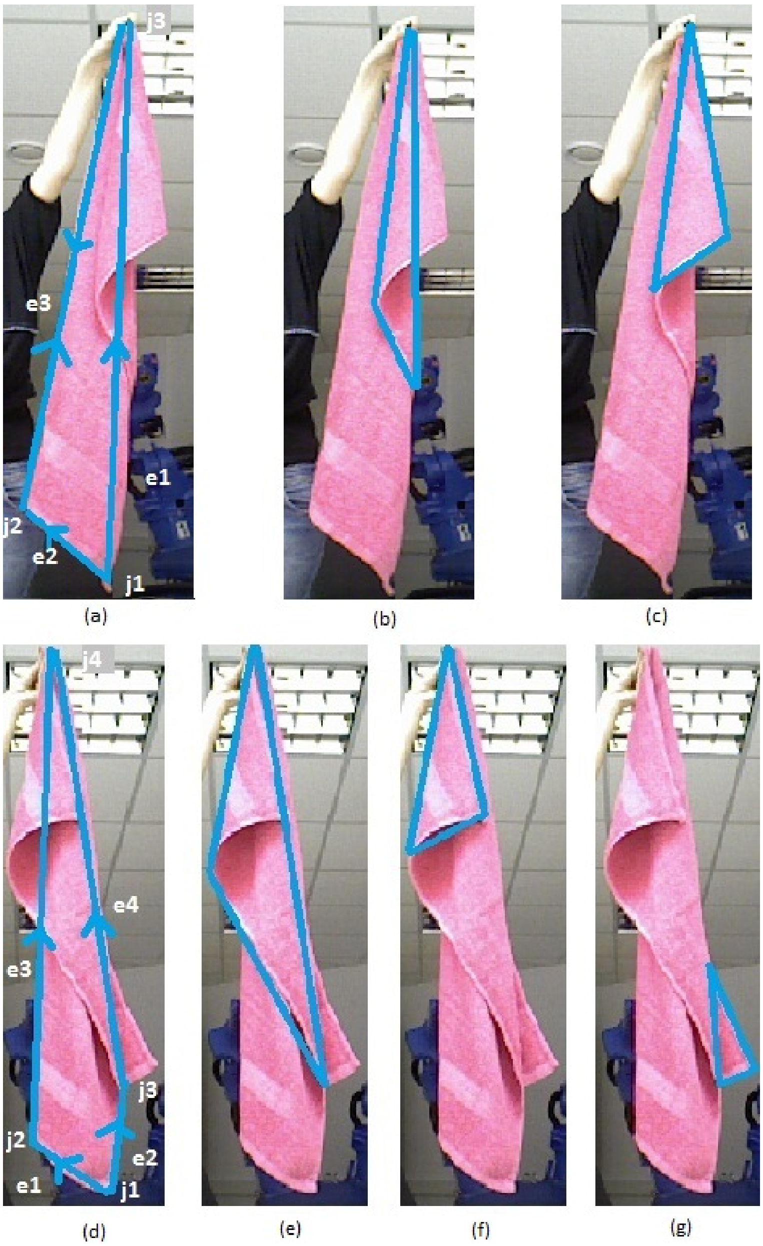


Figure 5: Detection of layers

## 3.3 Topological analysis

When the decomposition of the towel into layers is completed, their topological analysis begins. For each layer, its relative position to the other layers in an unfolded configuration of the towel is found. In addition, the layers' edges are mapped to the towel's outline edges when it is unfolded.

### 3.3.1 Problem description

The difficulty of the problem is that a towel, like any piece of clothing, is a highly deformable object with infinite degrees of freedom. So, although holding the towel from one of its corners reduces the number of configurations, the state of the towel cannot be predicted. In addition, the presented method uses data detected from only one side of the hanging towel. This means that there are hidden parts of the towel, in other words, hidden information about its configuration. Finally, another difficulty is that the layers are not flat to the surface of the camera, as a result, they cannot be used are parts of a puzzle that fit perfectly with each other and form an unfolded flat towel (even if it is with missing parts, as mentioned before).

The topological analysis is based on restrictions that are dictated by the towel's rectangular shape and observations concerning the configurations of a towel hanging from one of its corners. Actually, rules that determine the connection of the layers and indicators that suggest a preference to a candidate configuration are suggested. Firstly, the relative position of the layers with each other is defined and then their edges are mapped to the outline of an unfolded, straightened towel.

In the proposed method, the layer with the lowest hanging point is considered stable (as is mentioned in the next paragraph this layer includes one of the diagonals of the towel). Starting from the next layer with the lower corner, one by one, the rest of the layers are placed in the right or the left side of the already located layers (this applies for the layers starting from the grasping point; layers that start from a lower point are examined in the end). Depending on the side it is placed, the layer is left as it is or its symmetric is used. If there is a rule that bans one of the two possible positions or that dictates a definite connection of one layer with another, then there is only one solution for their connection. In cases where both positions are possible, there are indicators that suggest one of them without excluding the other one. When all the layers are taken into account, all the possible configurations are evaluated and, according to the indicators, one of them is the one that corresponds to the real configuration. Since the winning configuration is found, the edges of the layers are mapped to the sides of an unfolded

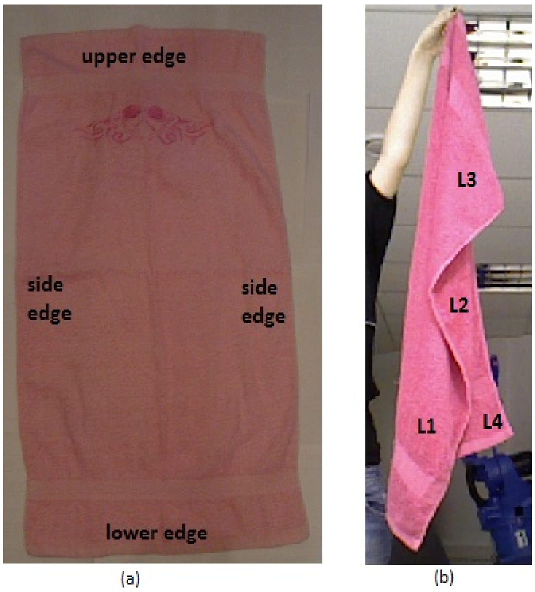


Figure 6: The towel in an unfolded and a hanging configuration

towel.

### 3.3.2 Rules for the relative topological connection of layers

The parts of the towel that are not visible to the camera and the different directions of the layers did not allow facing the problem of the hanging towel's topological analysis as a puzzle. So, rules based on the rectangular shape of the towel and on observations of the hanging towel's configurations are introduced in the following paragraphs. If the reader does not wish to read the details of these rules, he/she can skip this subsection and proceed to the following without any problems in the comprehension of the text.

The fact that the shape of a towel is rectangular dictates some restrictions about the possible topological relations between the layers. First of all, when a towel is held by one of its corners, then the lowest point of the hanging article corresponds to its not neighbouring corner. So, it is a priori known that the layer with the lowest point covers the diagonal of the towel (in figure 6.b it is layer L1). Nevertheless, it is not known if the edges of the lowest corner correspond to the lower or a side edge in an unfolded configuration(Fig 6.a). Another restriction that the towel imposes is that its shape is convex. So, the closest to the diagonal a layer is, the longer its edges, when it is compared with other layers placed on the same side of the diagonal. For example, in figure 6.b layer L3 cannot be placed closer to the diagonal than L2, or L4 cannot be left on the side of the layer L1 where it is currently located.

An important step for the topological analysis of the towel is the analysis of its current

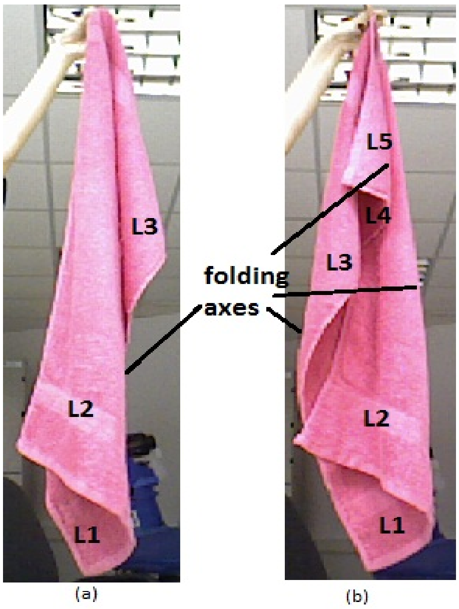


Figure 7: Connection of layers

hanging configuration. The connection of the layers is determined by their edges. Two layers might have one common edge, they might not have any edge in common or an edge of a layer might coincide with a part of the edge of another layer. In this case the layers are called neighbours (in fig 7.a L2 and L3 are neighbours, whereas in fig 7.b L3 and L4 are neighbours). The first case is rather interesting since its gives a definite topological connection between the two layers with the common edge. The two layers occur from a fold and correspond to neighbouring positions in the unfolded configuration of the towel (in fig 7.a L1 is connected to L2 whereas in fig 7.b L1 is connected to L2 and L3 and L4 is connected to L5). This common edge is called folding axis(Fig 7). In addition, if the folding axis is one of the edges that establishes two layers as neighbours in the hanging configuration, then the two layers cannot be at the same side of the towel's diagonal in an unfolded configuration ( in fig 7.a for layers L2 and L3). This occurs only for layers that start close to grasping point. Actually, if this was not the case, the layer with the folding axis had to extend behind the visible part of the towel and then continue to the neighbour layer. In this case, it would not be possible for the neighbour layer to start from a point close to the grasping point.

Finally, by observing the relations between layers, indicators about the most plausible relative position of two layers are introduced. These indicators are taken into account only when a definite connection between the layers has not been established by the previous rules. So, it is observed that a layer is usually topologically connected, in the unfolded configuration, to the layer and from the side of the layer with which their edges: 1) have

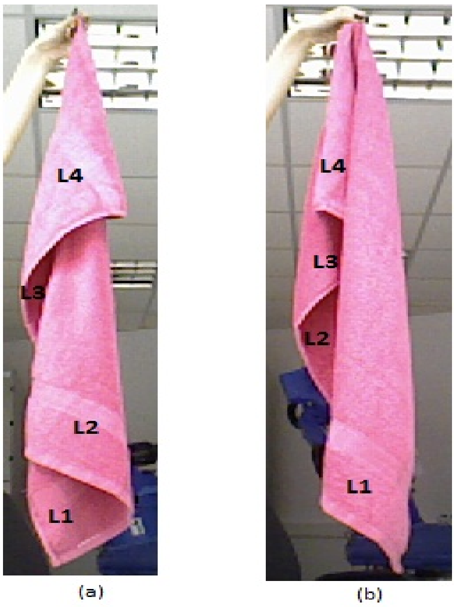


Figure 8: Configurations excluded from the indicators

the smaller difference in length, 2) are neighbours. As it is shown in fig 8.b, these conditions are valid for the layers L3 and L4. An exception is made and these indicators are not taken into account, in configurations like those depicted in figure 8. In these cases: 1) the layer with the diagonal is neighbour with another layer (in figure 8.b L1 is neighbour with L2) or 2) the layer, with which the layer with the diagonal is connected through a folding axis, completely overlaps it and is neighbour with another layer (in figure 8.a L2 completely overlaps L1 and is neighbour with L3). The exception is valid only for the connection of these neighbouring layers and not for the whole configuration.

### 3.3.3 Indicators for the correct final configuration

After the examination of the relative position that the layers could have in an unfolded state, more than one possible configuration might occur. Although there may be a preference for one of them according to the method's indicators, the whole configuration is evaluated in order to achieve better results. Once more, the rules suggested here cannot give a definite result on their own, but combined with the previous indicators they determine the solution. So, one indicator that a configuration is correct is that the smaller edge connected to the grasping point, which probably corresponds to the upper side of the towel, and the smaller edge connected to the towel's lowest hanging point, which probably corresponds to the lower side of the towel, incline towards the same side(Fig 9c). In addition, another indicator about the correct configuration of the towel is based on the lowest corner of the towel, when both of its edges are visible (e.g. when the layer which

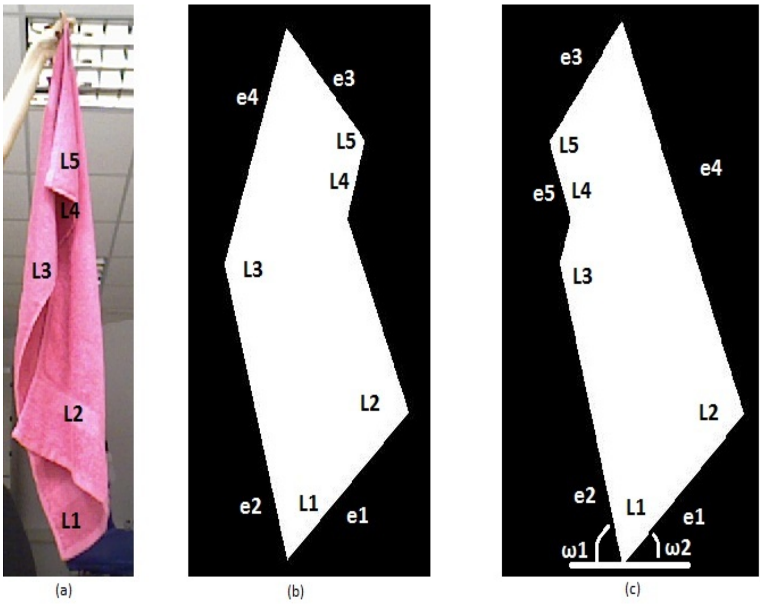


Figure 9: (a) the towel in a random configuration, (b) a candidate for the layers' topological connection, (c) the correct topological connection of the layers

includes this corner is rectangular, like L1 in 9a). Actually, the angle that the diagonal, which connects the grasping point with lowest hanging point, creates with the lower edge of the towel is bigger or, in the rare case of a square towel, even with the angle created by the diagonal and side edge of the towel (for convenience the complementary of these corners are calculated as shown in figure 9c). So, according to these angles the lower side of the towel can be found(e.g. edge e1 in Fig 9c ). If the smaller edge connected to the grasping point, which usually corresponds to the upper edge of the towel, inclines to the same side as the lower edge, then this constitutes an indicator that this configuration is correct.

In figure 9, two possible results from the analysis of hanging towel's state in 9a are shown. The shapes that occur provide the topological connection of the layers and not the actual shape of the towel. As it is obvious from 9a, in 9b the result is wrong, whereas in 9.c correct. The same conclusion occurs from the application of the two indicators for the correct final configuration. In figure 9c edges e1 and e3, which are the smaller edges connected to the lower point of the towel and the grasping point, incline towards the same direction. In addition, angle 2, is smaller than 1, implying that edge e1 is the lower edge of the towel.

## 3.4 Experiments

The presented method was tested with 50 different configurations of a hanging towel. Correct topological connection of the layers and correspondence with the edges of the towel in an unfolded configuration was achieved for 49/50 configurations, while for one case there was a tie between the correct result and a wrong one.

An example that proves the correctness of the method is shown in figure 9. The hanging towel is divided in 5 layers. For Layers L1, L2 and L3 and for L4 and L5, there is a definite topological connection through their folding axes. The fact that L3 and L4 are neighbours and that their edges have small difference in length indicate a possible topological connection between these two layers. Although there is a preference for the combination of the layers depicted in fig 9c, due to the relation between L3 and L4, the combination in fig 9b is not excluded. Both of them are evaluated for their final result. As has already been mentioned, the combination depicted in fig 9c satisfies the indications for a correct final topological connection. So, since all the indications agree with it, the topological analysis depicted in fig 9c is considered correct. In addition, according to the indicators, edge e1 corresponds to the lower edge of the towel, e4 to a side edge, e2 and e5 to a side edge other than e4. E3 corresponds to the upper edge of the towel or to an edge created by a fold close to it. The upper edge of a towel is not always visible to the camera (fig 7a), and since information about the texture of towel, which is different at the outline, is not used, it is not possible to be affirmative that an edge is the upper edge of a towel.

## 3.5 Conclusions

The presented method analyses the topology of a hanging towel. The towel is decomposed into layers, which are matched to positions in an unfolded configuration of the towel. The parts of the towel that are not visible to the camera and the different directions of the layers did not allow facing the problem as a puzzle. So, rules based on the rectangular shape of the towel and on observations of the hanging towel's configurations are introduced. Firstly, the relative position of the layers is examined and then the whole topological connection is evaluated. According to the final result, the edges of the layers are matched to the edges of the outline of the towel when it is unfolded.

The implementation of the method proved its correctness, since 49/50 configurations that where tested were analysed correctly, whereas for only one configuration (1/50) there were two possible topological combinations proposed, with one of them being the correct one. For these experiments the edges and their junctions were given manually in order to evaluate just the results of the layer extraction and the topological analysis.

The use of the topological analysis is that, in this way, good grasping candidates can be detected for the unfolding of the towel. Even in cases where a preferable grasping point, e.g. a neighbouring corner of the current grasping point, is not visible, the topological analysis provides information about the visible part that is closer to it. So, the appropriate manipulations can be preformed in order to make it visible to the camera and grasp it.

# 4 Clothes Template Matching

## 4.1 Introduction

Object recognition is a fundamental computer vision problem and, although significant progress has been accomplished in this field, it still remains a non trivial task. Recognition of articles of clothing, in particular, is considered even more challenging, due to their non rigid properties and their great variability in type, texture and style. In this work we are interested in recognition of the garments during their manipulation by a robotic system, in order not only to classify clothes according to their type, but to facilitate manipulation itself. Thus, the adopted recognition approach is based on template matching which in addition to identification provides also information about the configuration of the manipulated garments. A limited number of studies in the published literature address garment recognition in general. This number is even smaller for clothes manipulated by robots.

In Kaneko et al. [28] a planning strategy for isolating and unfolding clothes using dual manipulator is presented. According to that approach, after a grasping region is initially selected and the garment is picked up, two grasping points on the hemlines are used to rehandle it, in order to facilitate the classification task. Classification is performed in two stages, where at the first stage the type of the hanging garment's configuration is determined and at the second stage the appropriate geometric features are extracted and used for the final classification of the garment. The method discriminates between shirts, trousers and towels, presenting 90% classification accuracy.

A dual manipulator approach has been also employed in Osawa et al. [43] in order to unfold massive laundry, while classifying the manipulated garments to the appropriate types. A key feature of that approach is the lowest point manipulation, which results in specific configurations for the grasped clothes that can be represented by a small number of image templates. Then, classification is performed using simple template matching techniques, based on image covariance, yielding classification accuracy of over 90% when discriminating between 8 different types of garments.

In a different direction, Kita et al. [31] employed a trinocular stereo vision system to observe and recognize clothes during manipulation by a humanoid robot. In case the observation does not provide sufficient information for recognition, proper recognition-aid actions are automatically planned and executed. Recognition is based on fitting the acquired 3D shapes to deformable models of clothes generated through physical simulations as described in [33].

In Cusumano-Towner et al. [15], a Hidden Markov Model (HMM) is proposed for estimating the identity and configuration of garments manipulated by a two-armed robot. The transition and observation models of the HMM are based on cloth simulation. The observation model uses simple features, such as the height of the hanged garment when held by a single gripper and its silhouette when hanged by two grippers. The proposed model accurately estimates the identity and configuration of a large variety of manipulated garments, allowing the robotic system to autonomously bring them to a desired configuration.

Interactive perception is also employed in Willimon et al. [59], in order to aid garment recognition. The clothes are grasped by an arbitrary point and two images are acquired providing a frontal and a side view of the garment. The procedure is repeated ten times resulting to a total of 20 images for each garment. Four types of features are extracted from the acquired images based on area, eccentricity, binary edges, and Canny edges. The extracted features are employed by a nearest neighbour classifier, which discriminates between six types of garments. The presented results confirm the usefulness of the interaction stage, as the classification accuracy is increased by 59% by it.

A parametric approach to garment recognition is proposed in Miller et al. [40], according to which each clothing category is assigned with a parametrized shape model. Variation of the parameters accounts for the variety in shapes for clothes of the same category. The proposed approach extends previous work [39], where the parameters were fitted manually by a human operator. The extension is accomplished by a novel algorithm that when given an image of a garment is able to automatically find the parameters that provide the best fit to the model. The presented method allows garment recognition, simply by examining which model produces the best fit. However, a basic assumption of the method is that the garments are already crudely spread out on a flat surface.

## 4.2 Proposed Algorithms

In this work a template matching approach, similar to the one presented in [43], is proposed for garment recognition. The main difference is in the method selected for conducting the matching. In [43] simple template matching based on image covariance is performed, whereas we employ inner distance shape contexts [36] shape matching technique. The proposed method is extending the widely used shape contexts method [4] and presents useful properties that can benefit the garment recognition task. The employed shape matching technique is combined with the lowest point manipulation strategy presented in [43]. Thus, images of garments under lowest point manipulation are matched using inner distance shape context to a collection of selected prototypes. The class of the prototype presenting the closest match is assigned to the manipulated garment, whereas a correspondence between their contour points is established during matching.

*Inner distance shape contexts.* As argued in Ling et al. [36], part structure is very important for recognizing complex shapes, especially when these shapes include articulations, which allow for non linear transformations. Since articulations are present in most articles of clothing such as shirts, trousers, shorts etc., garment recognition has a lot to benefit if a shape descriptor that is insensitive to shape articulations is employed.

For this reason, the shape classification method proposed in [36], based on the concept of inner distance, has been employed in this work for the challenging task of recognizing clothes. Inner distance is defined as the length of the shortest path within the shape boundary and is considered to be sensitive to part structures but insensitive to shape articulations. A key advantage of the inner distance approach is that it is not dealing with part structure and articulations explicitly, avoiding this way ambiguities regarding partitioning of the shape.

Inner distance is a distance metric that can be combined with existing shape matching methods resulting in an extended version that is insensitive to the non linear transformations of the articulations. As proposed in [36] it can replace Euclidean distance metric when extracting shape context descriptors [4]. It should be noted, that shape contexts have been successfully employed for efficient visual search of garments in the content based image retrieval application presented in Tseng et al. [57].

However, the inner distance extension presents two basic differences with respect to the original shape contexts. The first difference is that instead of using only contour points to describe the shape, additional area information is employed. The second difference is that the ordering of the contour points is exploited by a fast dynamic programming algorithm [14], which performs silhouette matching. In Figure 10, a binary image of a shirt silhouette is illustrated, where the inner distance paths are depicted. In this figure, the red lines connecting contour points denote the inner distance paths that are the same with the corresponding Euclidean paths. The green line denotes a path that differs when the inner distance (right) is used instead of the Euclidean (left).

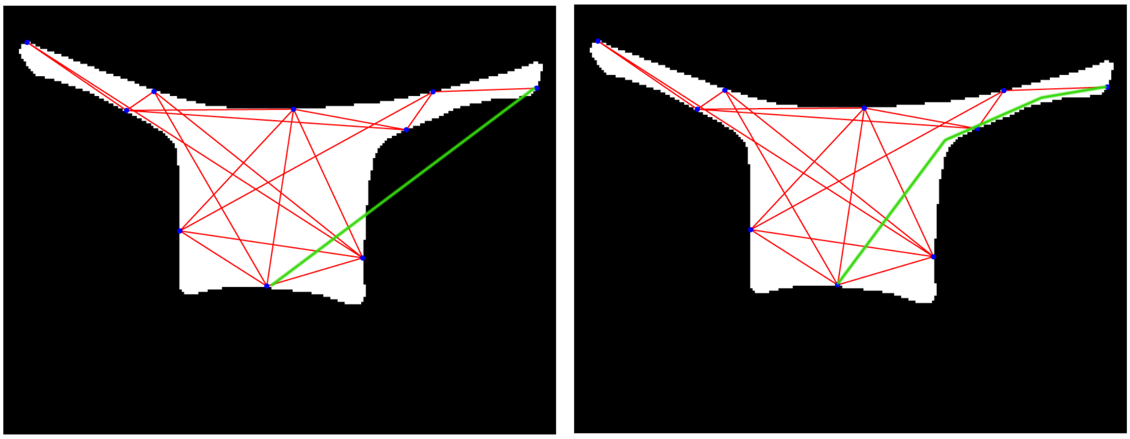


Figure 10: Binary silhouette of a hanging shirt. Red lines denote inner distance paths that are equivalent to corresponding Euclidean paths. Green lines denote an inner distance path (right) that is different to the corresponding Euclidean path (left).

The adopted method denoted henceforth as IDSC can be applied also to silhouettes with holes. However, only the outer boundary points are considered for matching. Hence, the inner distance between all pairs of these boundary points should be computed. Once these inner distances are computed, a polar histogram is constructed for every point, like in the original shape contexts method. Notice however, that if the shape is convex the inner distance is equivalent to the Euclidean and both methods produce identical descriptors. In Ling et al. [36], the concept of the inner angle is also introduced, which apart from scale and translation makes the proposed method rotation invariant, as well.

As argued in the aforementioned study, the proposed method is considered computationally efficient. The shortest path algorithm [14] employed for the calculation of the inner distances presents $O(n^3)$ complexity, where $n$ is the number of points. Then, the polar histogram is computed in $O(n^2)$ time. Finally, using the computed histogram and dynamic programming shape matching is conducted again in $O(n^2)$ time. Inner distance shape contexts performs matching between shape silhouettes and a matching cost is produced. Then, the similarity between shapes is assessed based on that cost value.

Apart from garment recognition, matching using inner distance shape contexts results in a correspondence between points on the matched contours. Hence, the contour of the matched prototype can be used to model the manipulated garment and the extracted model can be used for planning a strategy that brings the garment to a desired configuration. However, the details of such modelling and planning remain to be investigated in future work.

*Lowest point manipulation.* In order to use inner distance shape contexts for garment recognition, certain shape silhouettes should be selected as prototypes for the different

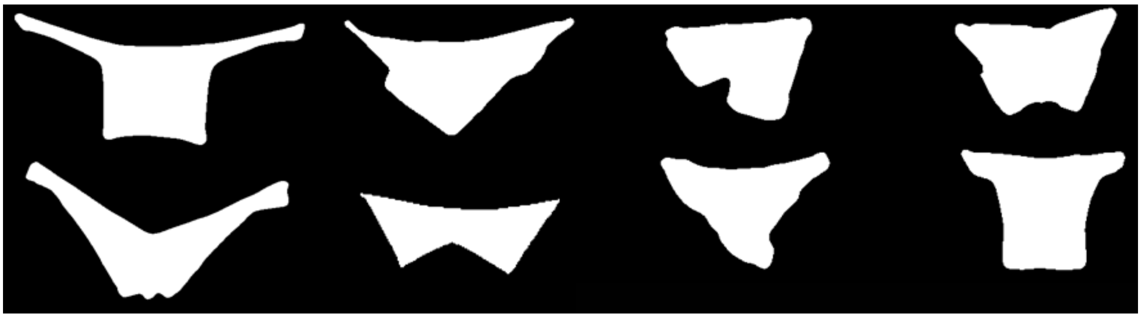


Figure 11: Binary silhouettes of the selected prototypes. The images have been extracted from real garments after lowest point manipulation by a human operator.

kinds of garments. Then, a query silhouette is classified to the garment category of the most similar prototype i.e. the one producing the minimum matching cost. In this work six types of garments are considered: Shirt, Trousers, Shorts, T-shirt, Skirt, Towel.

Since clothes are made of non-rigid textile fabrics that can assume a great number of different configurations, the selection of prototypes is challenging. For practical purposes the number of considered configurations should be reduced. One choice is to consider as a special case that the clothes are lying on a table flat and unfolded. However, knowing the type of the clothing article before actually unfolding can facilitate the unfolding task itself. Therefore, in this work a different configuration is considered. This configuration can be reached starting with an arbitrary initial configuration and performing the lowest point manipulation proposed in [43] and adopted in [15], as well. According to this approach the garment is hanged by an arbitrary part, the lowest point is identified and the distance L from the hanging point is calculated. Then, the lowest point is grasped and is used as the new hanging point. This procedure is repeated until convergence to the maximum L distance. After convergence, the lowest point is picked up and held near the hanging point, while the width between them is gradually expanded until it becomes equal to L distance. The proposed manipulation sequence converges to one or two possible configurations for each garment type. Hence, a reasonable number of prototypes can be selected, whereas the recognition can be performed before the unfolding task is completed. In Figure 11 the selected prototypes are illustrated. Except for the shorts and the T-shirt, which need two prototypes, the remaining types always result in a single configuration that can be associated with a single prototype. Therefore, only eight prototypes are sufficient for discriminating between these six garment types when the lowest point manipulation is performed.

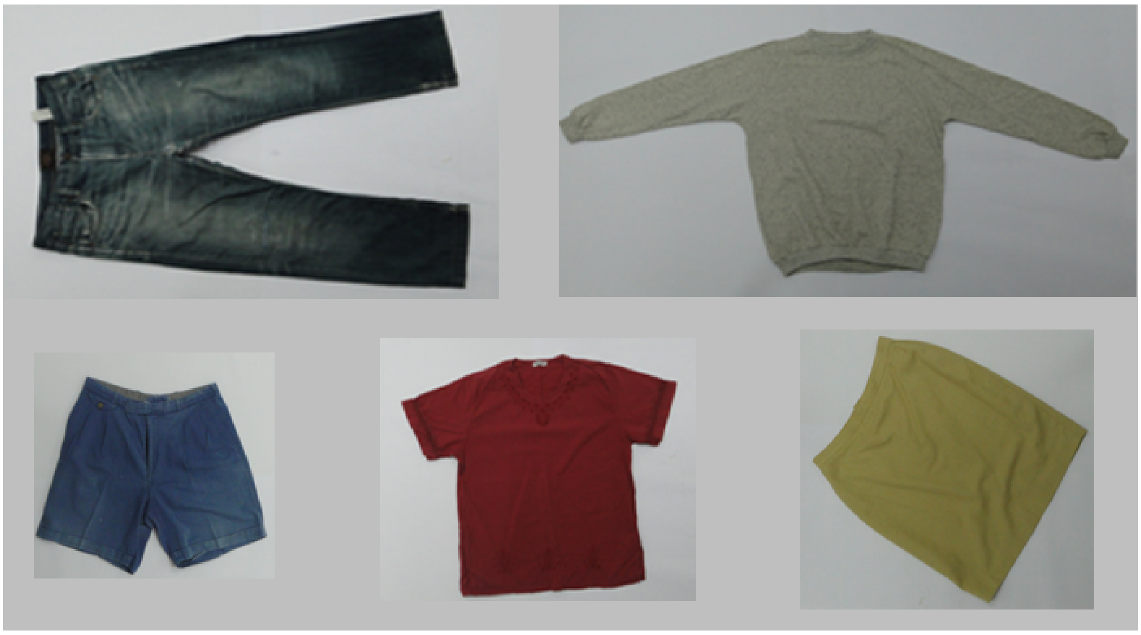


Figure 12: Typical garments used for the creation of the database.

## 4.3 Experimental evaluation

An image database of garments under lowest point manipulation has been created in order to evaluate the classification performance of the adopted approach of garment recognition. In Figure 12a collection of such garments before manipulation are illustrated as reference. A Sony DSC-HX1 camera (CMOS 9.1 megapixel) has been used for the acquisition. No flash or any other special lighting equipment was employed.

The database includes 52 images, eight of which were selected for the extraction of the prototypes presented in Figure 11, whereas the remaining 44 images are used for testing. In Figure 13 the eight images selected for the extraction of the prototypes are illustrated.

For every test image, inner distance shape context is applied and shape matching is performed between the test silhouette and each of the prototypes. Since IDSC outputs the matching cost between the two input silhouettes, the garment in the test image is classified to the class of the prototype presenting the minimum cost. When IDSC is employed, a classification accuracy of 95.5% is achieved. In Figure 14, the confusion matrix for the IDSC classification is presented. As shown in this matrix a T-shirt is misclassified to shorts, whereas a skirt is misclassified to a shirt. The remaining 42 items are correctly classified to their target class.

The performance of the original shape contexts method has been also evaluated for comparison. As expected, reduced classification accuracy (about 89%) is reported. These results indicate that the theoretical advantages of IDSC also apply in practice and garment

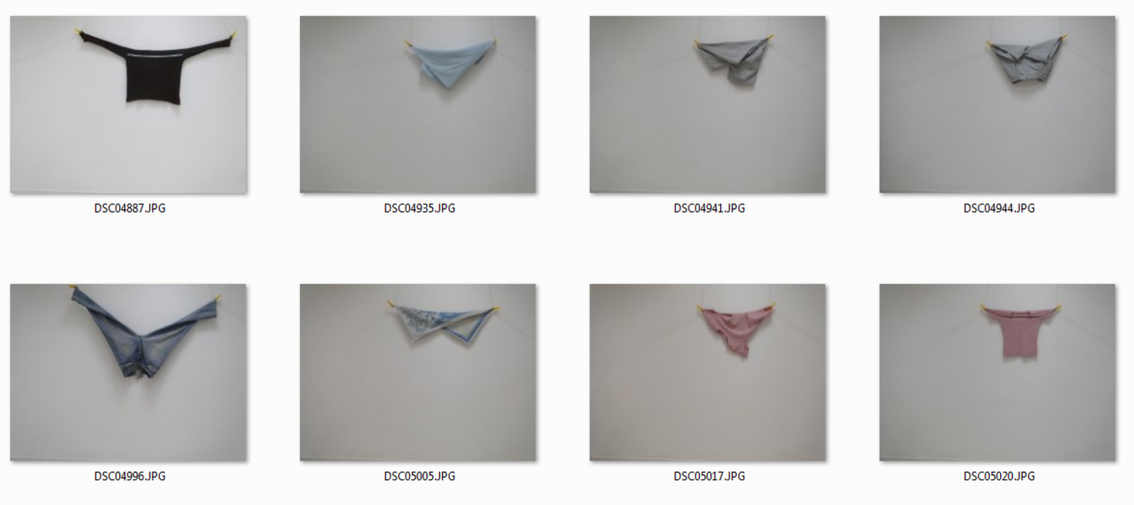
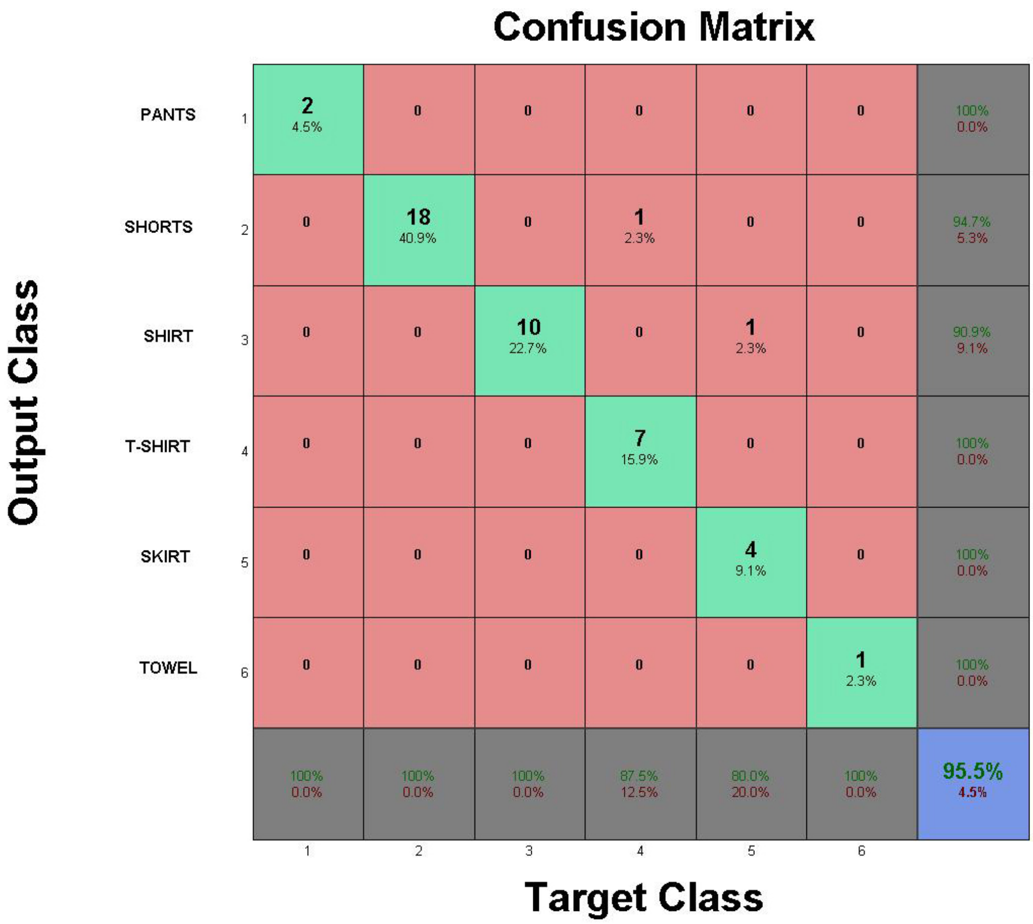


Figure 13: The original images of the selected prototypes hanged by the lowest points.



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| PANTS | 1 | 2<br>4.5% | 0 | 0 | 0 | 0 | 0 | 100%<br>0.0% |
| SHORTS | 2 | 0 | 18<br>40.9% | 0 | 1<br>2.3% | 0 | 0 | 94.7%<br>5.3% |
| SHIRT | 3 | 0 | 0 | 10<br>22.7% | 0 | 1<br>2.3% | 0 | 90.9%<br>9.1% |
| T-SHIRT | 4 | 0 | 0 | 0 | 7<br>15.9% | 0 | 0 | 100%<br>0.0% |
| SKIRT | 5 | 0 | 0 | 0 | 0 | 4<br>9.1% | 0 | 100%<br>0.0% |
| TOWEL | 6 | 0 | 0 | 0 | 0 | 0 | 1<br>2.3% | 100%<br>0.0% |
| | | 100%<br>0.0% | 100%<br>0.0% | 100%<br>0.0% | 87.5%<br>12.5% | 80.0%<br>20.0% | 100%<br>0.0% | 95.5%<br>4.5% |
| | | 1 | 2 | 3 | 4 | 5 | 6 | |

Figure 14: Confusion matrix of the garment classification based on inner distance matching

recognition benefits from the nice properties of the inner distance metric.

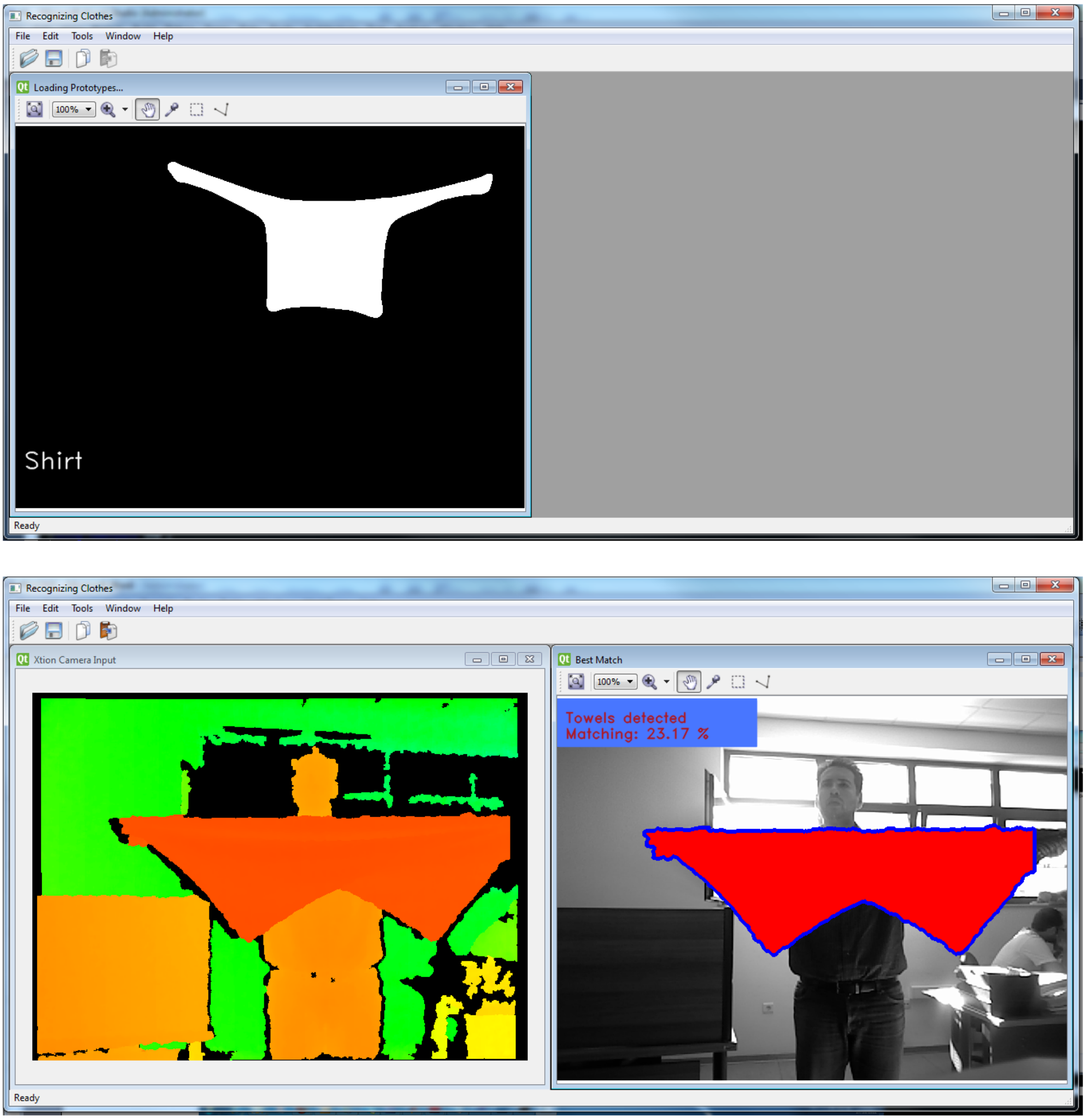
In order to perform recognition during manipulation, the proposed method has been implemented in C++ using also classes and methods of the OpenCV library, resulting in a real time application processing video frames. Moreover, a Graphical User Interface (GUI) has been designed in QT to facilitate the usage of the application. In order to perform fast and robust extraction of the garment silhouette an Asus XtionPro sensor has been employed providing both an RGB image and a depth map of the scene. The communication between the application and the Xtion camera is accomplished through methods of the OpenNI library. Figure 15 illustrates screen shots of the developed application, where a towel manipulated by a human operator is detected and recognized. The application has been tested with a large variety of manipulated garments and all cases resulted in successful recognition. For a more detailed description of the developed real time application refer to the corresponding section in D7.2.

## 4.4 Conclusion

In this work a method for visual recognition of clothes has been developed and a tool for real time garment recognition during manipulation has been implemented. The method is based on robust descriptors employed for shape matching and on a limited configuration space due to the performed lowest point manipulation of the clothes.

The method has been evaluated off line using images of real garments, whereas it was implemented and embedded to a real time system, processing video frames acquired on line by an Xtion camera. The presented results indicate that the aforementioned system is able to perform robust real time garment recognition, which is expected to facilitate autonomous manipulation of clothes by robotic arms.

However, during the recognition system's evaluation the lowest point manipulation has been performed by a human operator, who always succeeds to bring the garments to the desired configuration. Therefore, additional testing should be conducted after embedding the implemented system to the CloPeMa robot, in order to examine the applicability of the recognition method when the lowest point manipulation is performed by the robot. Furthermore, the performance of the recognition system should be tested in case a larger variety of garments and garment types is employed. Moreover, prototype selection should be further investigated in order to automatically select the most representative prototypes for each garment type.



Towels detected
Matching: 23.17 %

Figure 15: Screen shots of the developed garment recognition tool. On launch the prototypes are loaded (top). On XtionPRO activation two windows open (bottom), showing a depth map of the scene (left), and the detection and recognition results (right).

# 5 Modelling folded garments to facilitate unfolding

## 5.1 Introduction

Robot handling of non rigid materials, such as garments, remains an open research field. State-of-the-art of robotics indicates that a robot cannot manipulate such materials sufficiently. A challenging task regarding the manipulation of garments by robots is unfolding. A variety of research on unfolding articles of clothing using robots has been conducted so far.

In Hamajima et al. [21], unfolding is treated as a task of a more general handling process. In that study, the unfolding task is divided in three subtasks, i.e. hanging, rehandling, and shaping deformed parts. A similar approach to unfolding is adopted by Kaneko et al. [28], where a classifying task is added after the rehandling task. Classification during unfolding is also performed in Osawa et al. [43], which focuses on the unfolding of massive laundry. A significant difference from previous works is the proposed lowest point manipulation, which results in a limited number of configurations for the handled clothes.

In a different direction, Hata et al. [24] presented a robot system for cloth handling based on a 3-D vision system with stereo cameras and pattern projector. The system aims at the implementation of fully automated laundry factory, and it was tested on handling towels producing high success rate.

Maitin et al. [37] also address robotic handling of towels using stereo cameras. In that study, geometric cues are used for detecting grasp points for spreading out and the folding towels. The presented system produced a 100% success rate in 50 trials of picking up, spreading out and folding previously unseen towels.

A system able to handle different types of garments is presented in Cusumano et al. [15]. The garment's configuration and identity is estimated using a Hidden Markov Model (HMM) until it reaches a known configuration. The transition and observation models of the HMM are provided through cloth simulation. The estimated identity and configuration are used to plan the appropriate manipulations that bring the garment to the desired configuration.

In Willimon et al. [60], features such as corners, peak region and continuity of the cloth are used to determine a location and orientation in order to interact with the cloth and unfold / flatten it. The experimental results indicate that the proposed approach can significantly flatten out a piece of cloth.

In this work, a novel framework for unfolding clothes is proposed. The framework is composed by two main parts. The first one is modelling, where a model of the folded garment is extracted. The second one is planning, which uses the extracted model in order to construct an unfolding strategy. Although the presented approach is still under development and testing, preliminary results indicate the effectiveness and robustness of the methods.

This work focuses mainly on the modelling part of the framework. The garment to be modelled is crudely spread out on a flat surface and folded by an arbitrary axis. The key assumption of the proposed approach is that the folding axis becomes part of the outer boundary of the folded garment. This assumption is always valid when only a single fold exists. In case of multiple folds, the assumption holds for the last fold. The presented methods address the single fold case. However, the proposed framework can be extended to consider multiple folds as well.

An image of the folded garment is acquired and its outer boundary is extracted and approximated by a polygon. Then, each side of the polygon is examined and hypotheses are generated inferring the position and the orientation of the folding axis. In order to check the hypotheses, a prototype polygon corresponding to the unfolded garment is virtually folded and the resulted polygon is matched using inner distance shape contexts [36] to the original folded polygon. The hypothesis producing the best match is verified and the folded configuration of the garment is considered known.

The proposed framework based on hypotheses' testing and the inner distance metric is resilient to prototype selection. Thus, the employed prototype and the folded polygon may derive from images of different garments. Moreover, the proposed approach allows the use of multiple prototypes belonging to different types of garments, letting the selected hypothesis predict also the actual type of the folded garment. Namely, the clothes can be recognized while they are still folded and the predicted configuration can be exploited for planning the unfolding strategy.

## 5.2 Proposed Method

The modelling method of the proposed framework acts in two distinct phases. In the first phase hypotheses about the folding axis are generated, whereas in the second phase these hypotheses are tested using prototypes of unfolded garments. As previously described, the core assumption of this approach is that after folding, the folding axis becomes part of the outer boundary of the folded garment. Therefore, all sides of the outer polygon of a folded garment are considered to be potential folding axes.

An exhaustive search is performed examining all sides one by one. When a specific side is examined, the open contour that is formed by the remaining sides is matched to the contour of the prototype. Notice, that in order to keep the method robust, only outer contours that can be easily extracted are employed, whereas no edge information within the polygons is considered. However, matching the outer contours is extremely challenging, since the initial contour can be fragmented over several pieces after folding takes place. Moreover, some fragments can be missing due to overlaps, whereas fragments belonging to different parts of the initial contour can be wrongly connected after folding. The existence of fragmented correspondences indicates that only partial contour matching can be achieved.

In order to tackle the above challenges, the approach proposed in Riemenschneider et al. [47], has been adopted. In that study, a method performing object category localization by partially matching edge contours to a single shape prototype of the category is proposed. In the same study, a novel shape descriptor is employed, which is also adopted by our framework. A significant advantage of the adopted matching technique is that it enables efficient selection and aggregation of partial matches to identify and merge similar overlapping contours.

The adopted descriptor, denoted henceforth as delta descriptor, is calculated from the relative spatial orientations between lines connecting points sampled on the contours. Therefore, matching is highly dependent on the sampling distance, and in [47] this is compensated by searching over a large range of sampling scales. However, the adopted descriptors are translation and rotation invariant. When the sampling distance is fixed they are also scale invariant. Another important property of the adopted descriptor that is not identified in [47] is that it is, in some sense, reflection invariant. This property is very useful in case of folding, since the folding axis can be considered as a reflection axis for the unfolded contour. Hence, after folding some parts of the outer contour may actually represent reflections of parts belonging to the initial contour. A subtle point is that a single reflection inverts contour orientation, whereas delta descriptor is not invariant to such inversion. Therefore, in order for delta descriptor to be truly invariant to reflections, the contour orientation should be changed back to the original one. Since the outer contour of the folded polygon can be a collage of original and reflected parts of the unfolded contour, the following strategy is proposed.

An arbitrary contour orientation is selected for the unfolded contour and delta descriptors are extracted. Then, delta descriptors are also extracted by the folded contour using the original orientation. A second set of delta descriptors are extracted using again the folded contour but inverting its orientation. Then, matching is performed twice. The first set of descriptors is used for matching unfolded parts of the contours, whereas the second set is used for matching parts of the original contour to folded parts of the query contour. Each match is assigned an affinity score and after thresholding only strong matches remain. As described in [47] similar overlapping contours are merged and the remaining matches are aggregated.

In Figure 16a the simplified polygon corresponding to a folded T-shirt is illustrated, whereas the folding axis is denoted by a red dashed line. In Figure 16b an example of a partial match is illustrated. In this figure the prototype polygon, corresponding to the unfolded T-shirt, is depicted with black crosses, whereas the folded polygon is depicted with blue crosses. The part of the folded polygon denoted by magenta crosses is correctly matched to the part of the prototype polygon denoted by red crosses. The same applies for Figure 16c, only in this case both parts correspond to the unfolded region of the contour. Finally, in Figure 16d an erroneous match is presented.

In [47], the performed object category localization is also based on hypotheses. However, the methods employed in [47] for hypotheses generation and verification are not suitable for folded shapes and a different approach has been developed and integrated into our framework. Thus, in our approach every partial match generates an initial hypothesis about the position and orientation of the folding axis, with respect to the employed prototype. Then, a local affine transformation is estimated based on the matched parts and is applied to the examined side. When the resulting axis is outside the prototype contour, the hypothesis is dismissed. In fact, instead of using the initial contour, the prototype is eroded and a smaller contour is employed. This is done in order to discard erroneous hypotheses based on unfolded sides of the prototype, which after correct matching lead back to the correct side. However, due to imprecision, the prediction can be interpreted as a fold parallel to a side. Thus, the proposed approach does not cope with such folds if they are too close to the corresponding side. In case the initial hypothesis prediction is not entirely outside the eroded contour, it is preserved for further testing. The majority of the hypotheses are discarded after checking the position of the predicted axis. Therefore, fewer computations are needed in the final verification phase.

In Figure 16b an example of a generated initial hypothesis is illustrated. In this figure, blue circles denote the position of the fragment shown by magenta crosses after applying the affine transformation estimated by the match. The same transformation is applied to the examined folded polygons side denoted by the blue dashed line and the result is illustrated by the red dashed line. In the presented example, the generated initial hypothesis denoted by the red dashed line is actually valid. However, this is not yet verified and the generated hypothesis is preserved for further testing. In Figure 16c another valid hypothesis is generated, whereas in Figure 16d an invalid hypothesis is generated but since it predicts a legal position for the folding axis it is also preserved for further testing.

Different partial matches can result in very similar hypotheses. This is expected in

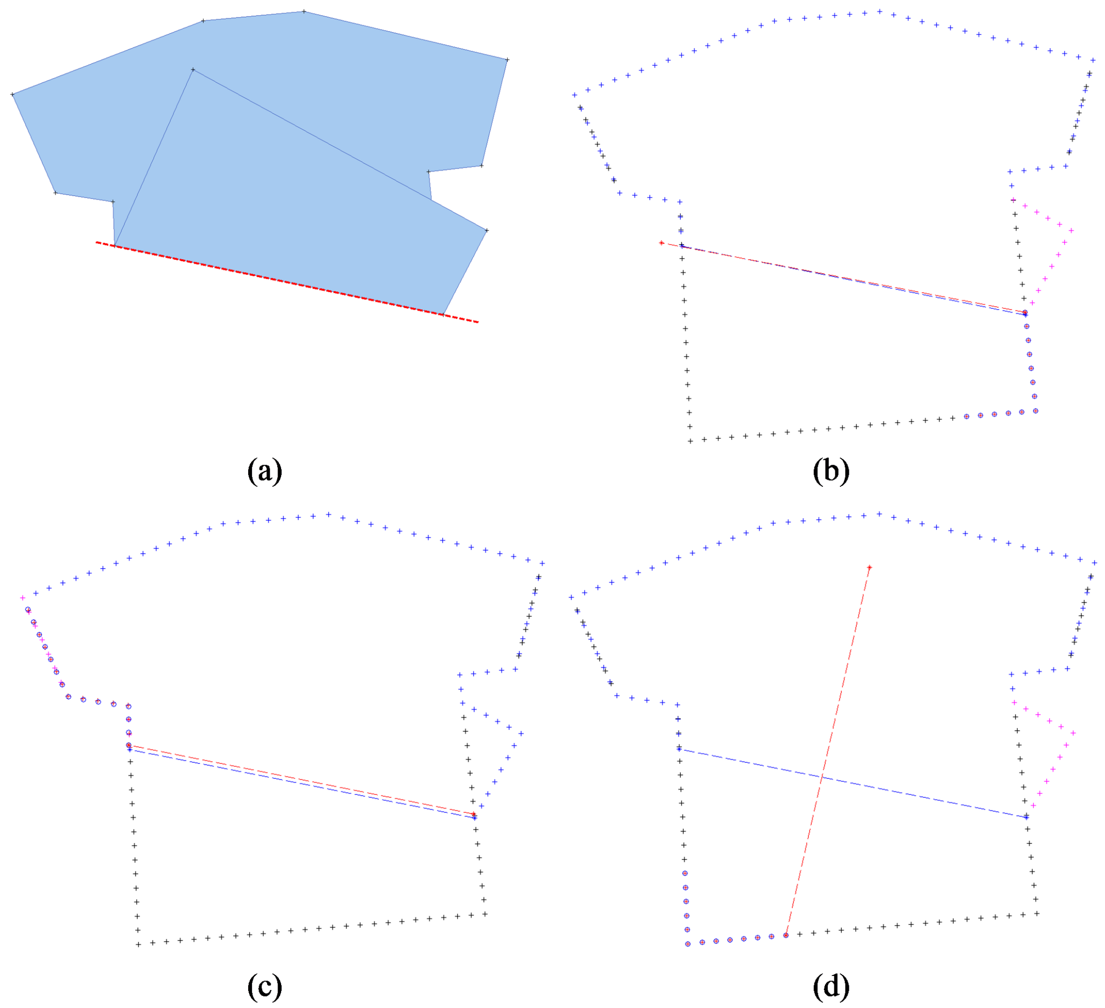


Figure 16: A virtually folded T-shirt, where the red line in (a) denotes the location of the actual fold. The red dashed lines in the remaining subfigures denote the predicted locations of the fold according to the generated hypotheses.

case the matches are correct and it is exploited by aggregating similar initial hypotheses before proceeding to further testing. Each initial hypothesis is assigned with a confidence score $Cs$, which is computed as a product of two other scores. The first one is the affinity score $As$ of the match used for generating the hypothesis. The second score $Raxis$ is given by:

$$Raxis = \begin{cases} e^{-\frac{Lout}{Lin}}, & \text{if } Lin \neq 0 \\ 0, & \text{if } Lin = 0 \end{cases} \tag{2}$$

where $Lout$ denotes the axis length that is predicted to be outside the prototype and $Lin$ denotes the axis length that is predicted to be inside the prototype.

The higher the values of $As$ and $Raxis$, the greater the confidence on the validity of a certain hypothesis. Of course, as already explained in case of $Raxis = 0$, the hypothesis is dismissed in an early stage.

The remaining hypotheses are aggregated in a similar manner to the one used for the selection of the matched fragments. Namely, a binary image is calculated, where each pixel represents a hypothesis and connected components analysis is performed. In this case the horizontal image axis corresponds to angles, the vertical to distances and each pixel represents the folding axis of each hypothesis, given in polar form. Therefore, similar hypotheses are expected to create connected regions in this binary image. After identifying a component, the associated confidence scores are summed and the result is thresholded. In the case when the components hypotheses are not dismissed, they are merged in a single hypothesis by calculating a weighted average of the involved axis. The corresponding confidence scores are used as weights for the calculation.

As a result of the above procedure, only a small portion of the initial hypotheses survive. Only these hypotheses are tested in the final verification stage, which is perhaps the most computationally intensive.

Each hypothesis determines a folding axis on the prototype contour. The predicted axis is used to virtually fold the prototype and a predicted folded contour is generated. The predicted contour is matched to the folded contour using inner distance shape contexts and a matching cost is estimated. The hypothesis resulting in the predicted contour that presents the minimum cost is selected as the most probable one. In case the estimated cost is lower than a predefined threshold, the selected hypothesis is verified. The matched virtual contour corresponding to the selected hypothesis is used to model the folded garment. The extracted model provides the location of the folding axis and assigns contour fragments to different folded parts.

The estimated model can be used for planning the unfolding strategy. However, some additional visual cues are needed, since the model is agnostic to which side is on top. The exact planning strategy is still under investigation. The same applies for the extension of the modelling method in order to deal with garments presenting multiple folds. In principle, a modification of the final verification stage combining different hypotheses should be able to predict contours under multiple folds. However, the effectiveness of such approach remains to be experimentally verified, since testing even a small set of hypotheses may become computationally intractable.

## 5.3 Experimental Evaluation

A series of experiments has been designed in order to assess the effectiveness of the proposed approach. As described above the adopted matching technique is sensitive to the selection of the sampling distance. Hence, in the conducted experiments a fixed sampling distance is employed, minimizing any mismatches due to discrepancies between point correspondences.

The first set of experiments examines the ideal case, where a prototype is virtually folded by an arbitrary axis and the proposed approach is applied to the resulting contour, using for reference the original prototype. This allows for extensive testing employing a large number of different folding axis, whose location on the unfolded contour is known. Since folding is performed virtually, only rigid transformations are considered. This implies that once matching is correct, the locally affine assumption always holds. The robustness of the proposed approach in case of non-rigid transformations remains to be examined in future experiments using images of garments that are physically folded.

Five different types of garments have been considered. Namely, the examined types include T-shirts, shirts, trousers, shorts and skirts. Real garments have been employed and the unfolded contours were extracted sampled by roughly 120 points. For each garment 60 different folding axes have been arbitrarily selected using pairs of contour points at random. Then, the garments have been virtually folded and the derived contours were resampled using the same sampling distance as for the prototype contour. An example of the detection experiments is illustrated in Figure 17 for different prototypes and folding axes. The images of the prototypes are depicted on the left hand side and the selected folded axes are denoted by red lines. In the right hand side the detected axes and the corresponding virtual folds are presented.

The main measure used for the evaluation of the prediction results is the difference between the actual and the predicted position and orientation of the folding axis. The position and orientation of the axis is provided in polar form and is defined by only two parameters. The first one is the distance $\rho$ of the axis from the origin of the selected coordination system, whereas the second parameter is the angle $\theta$ that the folding axis forms with respect to the $x$ axis. The distance is originally measured in pixels but for facilitating the interpretation of the results it is normalized with respect to the major axis of the ellipse that has the same normalized second central moments as the region defined by each prototype.

A summary of the results is provided in Table 1, where the mean differences $d\rho, d\theta$ of the aforementioned parameters over the 60 folds and the corresponding standard deviations are provided for the examined garments, after discarding outliers. The presented results provide a quantitative assessment about the accuracy of the proposed approach. However, there is not a clear threshold between a correct prediction and an erroneous one. A 0.1 difference in the normalized distances and a 5 degrees difference in the angles have been subjectively selected as a reasonable tolerance of the predicted axis deviation from its actual position and orientation, respectively. In fact, these values were also used for determining the outliers before calculating the aforementioned statistics. In the third column of Table 1 the correct detection rate based on the defined tolerance is given for each garment type.

The highest detection rate is presented by the T-shirt and the Skirt, where only one case resulted in detection failure. The lowest detection rate is presented by the Shirt, where three folding scenarios resulted to detection failures. However, there is also a qualitative difference between the different cases. In some of the failures the problem is only in the accuracy of the detection, due to imprecision in the estimation of the employed affine transformations. On the other hand there are cases where the failure is caused by verifying completely erroneous hypothesis based on wrong matches. In these cases we can objectively assess that the method is failing. Thus, in the last column of Table 1 the absolute number of failures caused by verification of erroneous hypotheses based on mismatches is presented.

In the total of 300 folding scenarios only 4 cases resulted in verifying erroneous hypotheses. However, there are 5 more cases, where detection fails because the detection accuracy is above the specified tolerance. The remaining 291 folding scenarios result in accurate axis detection indicating the robustness of the proposed approach. In general, similar results are produced by the different garment types. However, in case of Pants, the proposed approach presents slightly inferior performance with respect to accuracy and hypothesis verification.

The next set of the experiments evaluates the methods performance when the garments that are used for folding are different from the prototypes used to predict the folding axis position. Although this phase of testing is not complete yet, initial results based on subjective evaluation of the predicted folds indicate that the method is robust to moderate

Table 1: Folding axis detection results for 5 garments, using 60 folding scenarios for each

| Type | dρ | dθ | Detection Rate | Mismatches |
|---|---|---|---|---|
| Pants | $0.019 \pm 0.020$ | $1.75 \pm 1.82$ | 96.7% | 2 |
| Shorts | $0.014 \pm 0.019$ | $0.96 \pm 1.01$ | 96.7% | 1 |
| Shirt | $0.017 \pm 0.018$ | $1.17 \pm 1.04$ | 95% | 1 |
| T-shirt | $0.016 \pm 0.022$ | $0.90 \pm 1.12$ | 98.3% | 0 |
| Skirt | $0.014 \pm 0.015$ | $1.04 \pm 0.91$ | 98.3% | 0 |

differences between the prototypes and the unfolded contours of the folded garments.

Initial testing using multiple prototypes of different types for detecting the same fold indicates that the proposed method is able to discriminate between different garment types based on the prototype that generates the hypotheses with the minimum final matching cost.

## 5.4 Conclusion

In this work a method for modelling folded garments in order to facilitate unfolding has been proposed. The method is based on the simple but reasonable assumption that the folding axis becomes part of the outer boundary of the folded garment. Hence, each side of the folded contour is considered to be potential folding axis. Using unfolded prototypes, hypotheses can be generated about the position and orientation of the axis with respect to these prototypes. The hypotheses are generated based on partial matching between the prototypes contour and the folded contour. Using only contour information is an important attribute of the proposed method, since when it is applied to real garments it is not affected by their great variation in texture and colour.

The proposed approach to garment unfolding is a work in progress. However, the presented quantitative evaluation documented high accuracy in folding axis detection. Moreover, the correct detection rate for the 300 virtual folds that were tested is over 97%. The presented results of these initial experiments indicate that the proposed method could be employed for modelling a folded garment before planning the unfolding strategy.

As a next step, the methods tolerance to deformations should be quantitatively assessed. In order to do that, clothes should be folded both virtually and physically and the results should be compared. Although the method is designed to be insensitive to mild deformation, in case the comparison results are not satisfactory, the method should be modified in order to explicitly cope with deformed contours.

The selection of prototypes also needs further investigation. So does the impact of

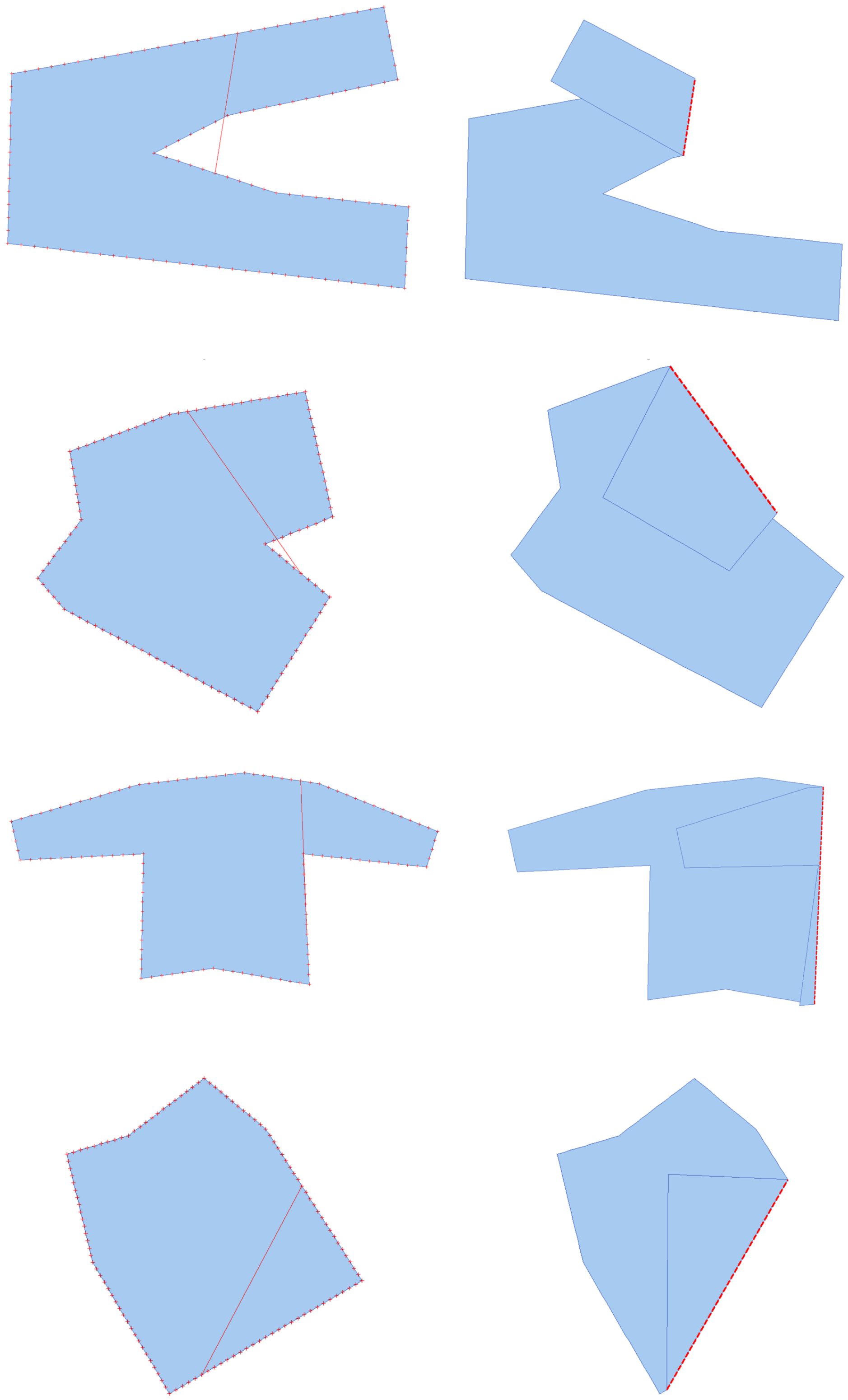


Figure 17: A virtually folded T-shirt, where the red line in (a) denotes the location of the actual fold. The red dashed lines in the remaining subfigures denote the predicted locations of the fold according to the generated hypotheses.

increasing the number of prototypes on the performance and reliability of the system.

An image database of unfolded garments should be acquired and a strategy for selecting prototypes for each garment type should be developed. A popular selection approach is using k-medoids clustering [30], whereas the number of clusters can be automatically selected using the silhouette method [49]. Then, the discriminating ability of the method between different prototypes should be experimentally evaluated. Finally, folding scenarios that include multiple folds should be considered and appropriate modifications should be made to the presented method for modelling folded garments.

However, even if the method is applied effectively and the model of the folded garment is successfully extracted, the derivation of the remaining steps for the actual unfolding is not straightforward. An initial direction that will be examined for the planning part of the framework is the geometric approach presented in Miller et al. [39, 40], which in that study is employed for the folding task.

# 6 A Framework for High-level Understanding

One of the key challenges of clothes recognition and perception is handling the variabilities in geometry and appearance. This variability is due to the large number of different configurations that clothes can take, self-occlusion and the wide range of fabric textures and colors. On the other hand, clothes are characterized by a consistent high level structure that may be represented concisely by a small number of production rules upon the clothes parts. And-Or graph grammars are the ideal architecture to represent an object by its parts which are related by specific rules. Moreover, graph grammars can handle situations of self-occlusion as they can recognise an object from a small number of pieces of evidence. In the following subsections the basic theory behind And-Or graph grammars is introduced and adapted to the problem of clothes understanding.

## 6.1 State-of-the-Art

Textiles do not have a stable shape and cannot be manipulated on the basis of an a priori geometric knowledge. However, the solution of the “laundry problem” is strictly related to the successful perception and manipulation of clothes, as described in the CloPeMa DoW. As has been described in Section 2.5 of the D1.1 deliverable, the most common approach to address this problem is a sequence of elementary manipulation actions [59, 44, 28]. All these approaches grasp the garment sequentially from different points of interest, until enough information becomes available to permit garment classification.

Within the CloPeMa project this problem will be encountered through the application of a combination of logic rules of perception that model human reasoning. As will be shown in the following sections, eventhough the development of this architecture is still at an early stage, the preliminary results seem very promising for the successful solution of the perception problem. The core of this architecture is the well known "Graph Grammars" [63], which include a tree structured decomposition of the contents of a scene, from scene labels to objects, parts and primitives, and a number of spatial and/or functional relations between the various tree nodes and for every level of the tree.

The application of Graph Grammars to scene interpretation dates back to the 70s [42], but it was just a few years ago that this architecture received intense scientific interest, partly due to the limitations of the appearance based and machine learning methods when they are scaled up. Despite the various implementation approaches reported in vision literature, most of them share a common tree architecture. They are represented as an And-Or graph, where each Or-node defines possible alternative sub-configurations, and each And-node merges a number of lower level components to a more complex higher level component.

To our knowledge there exist no application of such architecture to confront the clothes classification as defined by the "laundry problem". The current state-of-the-art of this approach, in regards to garment perception, is limited to clothes modelling and recognition, applicable to vision problems and graphics tasks, such as dressed human recognition and tracking, human sketch and portrait.

One of the most noteworthy representatives in this field is the work of Chen H. *et al.* [10]. In their 2006 publication they described a context sensitive grammar in an And-Or graph representation, which produced a large set of composite graphical templates to account for the wide variabilities of clothes configurations, such as T-shirts, jackets, etc. In a supervised learning phase, they asked an artist to draw sketches on a set of dressed people, and they decomposed the sketches into categories of clothes and body components: collars, shoulders, cuff, hands, trousers, shoes etc. Each component had a number of distinct sub-templates (sub-graphs). These sub-templates served as leaf-nodes in a big And-Or graph where an And-node represented a decomposition of the graph into sub-configurations with Markov relations for context and constraints (soft or hard), and an Or-node was a switch for choosing one out of a set of alternative And-nodes (sub-configurations) - similar to a node in stochastic context free grammar (SCFG). This representation integrated the SCFG for structural variability and the Markov (graphical) model for context. An algorithm which integrated the bottom-up proposals and the top-down information was also proposed to infer the composite clothes template from the image.

Graph grammars are more widespread for applications concerning human pose estimation. Zhu L *et al.* in 2008 [62] presented a novel structure learning method, Max Margin And-Or graph (MM-AoG), for parsing the human body into parts and recovering their poses. Their method represented the human body and its parts by an And-Or graph, which is a multi-level mixture of Markov random fields (MRFs). Max-margin learning, which is a generalization of the training algorithm for support vector machines (SVMs), was used to learn the parameters of the And-Or graph model discriminatively.

In 2011 Wu T. *et al.* [61] published a numerical study of the bottom-up and top-down inference processes in And-Or graphs using human faces as high-level vision examples. Also in 2011 Rothrock B. and Zhu S.C. proposed a stochastic grammar model that represented the body as an articulated assembly of compositional and reconfigurable parts [48]. The reconfigurable aspect allowed a compatible part to be substituted with an alternative part with different attributes, such as for clothing appearance or viewpoint foreshortening. Relations within the grammar enforced consistency between part attributes as well as geometry, allowing a richer set of appearance and geometry constraints over conventional articulated models.

Finally Morariu V.I. *et al.* in 2012 described a framework that leveraged mixed probabilistic and deterministic networks and their And-Or search space to efficiently find and track the hands and feet of multiple interacting humans in 2D from a single camera view [41]. Their framework detected and tracked multiple peoples heads, hands, and feet through partial or full occlusion. It required few constraints (did not require multiple views, high image resolution, knowledge of performed activities, or large training sets); and made use of constraints and And-Or Branch-and-Bound with lazy evaluation and carefully computed bounds to efficiently solve the complex network that resulted from the consideration of inter-person occlusion.

Graph grammars have been applied to numerous other applications, from room scenes classification to object recognition [22, 35, 53, 34]. One important feature that is exploited in many of these applications is the fact that graph grammars can provide reasonable accuracy in classification even with large portions of the inspected objects under occlusion. References [41] and [61] are exceptional examples of this functionality. Moreover, [63] is an excellent reference for numerous case studies where graph grammars are used not only for scene interpretation in the presence of occlusions, but also for prediction-through-training of the occluded components.

## 6.2 Graph Grammars Basic Theory

Graph grammars, as introduced in [10], can be represented by a 5-tuple as:

$$G = (N, T, R, \Sigma, A)$$

where:

- $N$, are the non terminal nodes. They consist of the *And*-nodes $U$ and the *Or*-nodes $V$. So we can say that $N = U \cup V$. *And*-nodes are configurations which consist of a number of sub-configurations, according to some production rules $r_1, ..., r_k \in R$, ($k$ denotes the number of rules). *Or*-nodes are configurations which can alternate between different sub-configurations.
- $T$ are the terminal nodes or the primitives. Primitives are the smallest possible configurations that can be detected in the image. They play the role of the letters in text graph grammars. They are the elements from which words are constructed.
- $R$ is the set of rules. These rules are used in order to describe how a configuration can be composed by smaller sub-configurations. Rules also contain the spatial relations and constraints between the sub-configurations. For example, a rule can describe how trousers are composed of two legs and how they are arranged geometrically in order to form the trousers.
- $\Sigma$ is the set of valid configurations, i.e. the different configurations the initial image can take when parsing it with the graph grammar. In the case of clothes recognition, the configurations are the clothes categories such as trousers, shirts and skirts.
- $A$ is a set of attributes that are relevant to each node or sub-configuration in the graph grammar. For example, the trousers can have as attributes the size of the legs and the length of the waist opening.

A simple graph grammar is shown in Figure 18. The squares correspond to the terminal nodes, i.e. the primitives. The first layer above the terminal nodes, nodes E, F, G, H, are the first sub-configurations consisting of primitives. The graph can grow up to the root node which can describe a whole scene.

## 6.3 Boundary simplification and primitives extraction

Clothes in general can contain any kind of textures or colors. This makes the interior of clothes less informative about their category. At this stage in our research we consider

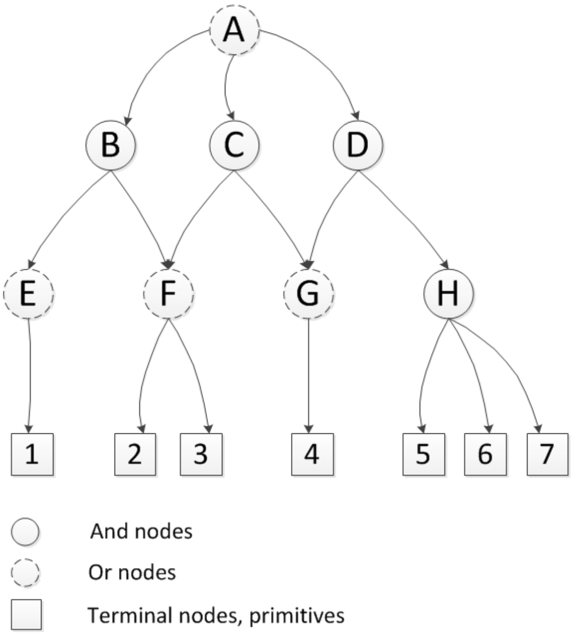


Figure 18: An And-Or graph grammar.

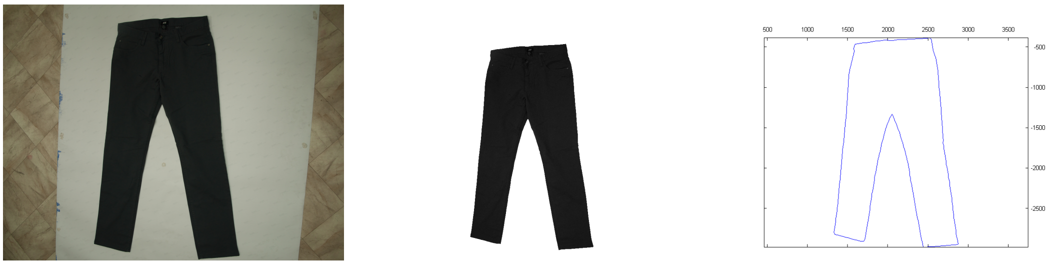
only the shape of clothes and especially their boundaries. Furthermore, for simplicity, we assume that a garment is lying on a table and most of its parts are unfolded.

The image acquisition is implemented under a uniform, white background, in order to extract the clothes through background subtraction. This process will be refined when the robot installation is finalized and the image acquisition system is fully functional. In Figures 19(a) and 19(b) an example of this process is presented.

Having segmented the garment from its background, we then trace its boundary points. The Douglas-Peucker Polyline Simplification [25] is then applied to these boundary points in order to get a simple sketch of the boundary of the garment, as shown in Figure 19. This sketch, composed of 300 boundary points, represents the input to the primitives extraction procedure.

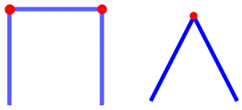
In its current implementation, the developed architecture confronts the perception problem assuming that there are only two kinds of garments available, trousers and shirts. The primitives have to be small boundary segments, which are structural elements for the boundaries of both trousers and shirts. We introduce two such primitives, which are symbolically called “Π” and “Λ” for their structure, as shown in Figure 20.

These primitives are extracted by tracing the simplified boundary using heuristic similarity measures, which are scale and rotation invariant. After the boundary tracing, a number of basic primitives are collected and form the terminal nodes of the graph grammar. These primitives will construct more complex configurations at higher levels of the



(a) (b) (c)

Figure 19: a) Original image. b) Trousers segmented from the background. c) Trousers' boundary after simplification.



(a) (b)

Figure 20: a) Primitive "Π". b) Primitive "Λ". Corners are marked with red.

tree using the rules and the structure of the graph grammar.

## 6.4 Logic Rules and Attributes

In this section we will introduce the logic rules used in our graph grammar along with their node attributes. Having the primitives extracted, we can apply the rules in order to form more complex configurations, until we finally reach a definite garment configuration. This procedure is called *bottom-up parsing* or *bottom-up detection*.

### Logic Rules

The first 3 rules are directly related to the primitives. They can be expressed as:

- *rule 1:* A "Π" is above a "Λ" and they have the same direction.
- *rule 2:* A "Π" is at the bottom right of a "Λ" and they have opposite directions.
- *rule 3:* A "Π" is at the bottom left of a "Λ" and they have opposite directions.

These rules are illustrated in Figure 21(a) to 21(c) where the direction of the primitives is also defined and their application to trousers is shown in Figures 21(d) to 21(f). For each rule applied, we construct a new node on our graph which has as children the primitives of the rule. Thus a connection is created on the graph between the children and the node.

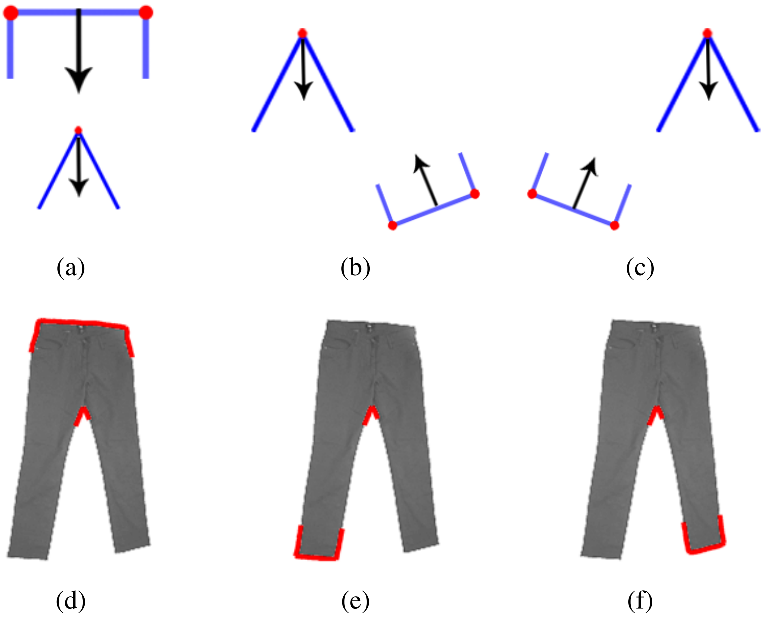


Figure 21: a-c) Rules 1-3. d-f) Rules 1-3 examples on trousers.

The new node contains a more complex configuration, consisting of the combination of two primitives. This configuration is denoted as $config^r$, where $r$ is the rule number that created the node. Nodes that have been created according to rules 1-3 form the second layer of the graph, just above primitives. Figure 22 shows how primitives can be combined to form each rule and how they are adopted to the graph.

Working similarly, we can produce the third layer of the graph and so on. This procedure of constructing the graph grammar based on our initial image and the primitives currently extracted is known as *parsing on the fly*. This means that we do not have a static graph grammar that is used for the whole database of clothes but a new graph grammar is constructed on the run for each new garment that our system is currently processing.

The third layer of nodes in our graph grammar is created using a new set of rules, this time by using only the nodes of the second layer as structural elements, as shown in Figure 22. This set of rules are defined as:

- *rule 4:* A $config^1$ appears together with a $config^2$ and they share the same “Λ” primitive.
- *rule 5:* A $config^1$ appears together with a $config^3$ and they share the same “Λ” primitive.
- *rule 6:* A $config^2$ appears together with a $config^3$ and they share the same “Λ” primitive.

- *rule 7:* A $config^2$ appears together with a $config^3$ and they share the same "Π" primitive.

As expected, moving higher in the graph hierarchy the configurations begin to resemble better the shapes of the specific clothes. More specifically, configurations 4-6 are getting the shape of the trousers, while configuration 7 is getting the shape of a shirt, as can be observed in Figure 22.

The last step in the bottom-up parsing procedure is to combine configurations 4-7 to form the 4th graph layer. The nodes of this layer contain full configurations of trousers or shirts. The last rules are:

- *rule 'Trousers'*:

  A $config^4$ appears together with a $config^5$ and they share the same $config^1$.

  or

  A $config^4$ appears together with a $config^6$ and they share the same $config^2$.

  or

  A $config^5$ appears together with a $config^6$ and they share the same $config^3$.

- *rule 'Shirt'*:

  A $config^6$ appears together with a $config^6$ and the $config^2$ of the first one is the $config^3$ of the second one.

  or

  A $config^7$ appears together with $config^2$ and $config^3$ sharing one "Λ" with the $config^2$ and the other "Λ" with $config^3$.

The final nodes created by these rules are connected to the root node. The root node is an *Or-node* which can lead to one of the various clothes configurations.

### Attributes

In order to evaluate how the rules correspond to clothes configurations, we assigned some attributes to each node. These attributes mainly hold distance units between primitives, but in future they will contain various kinds of features. In our graph grammar, $A$ can be written as:

$$A = \{a^i_j : i \in R, j = 1..M\}$$

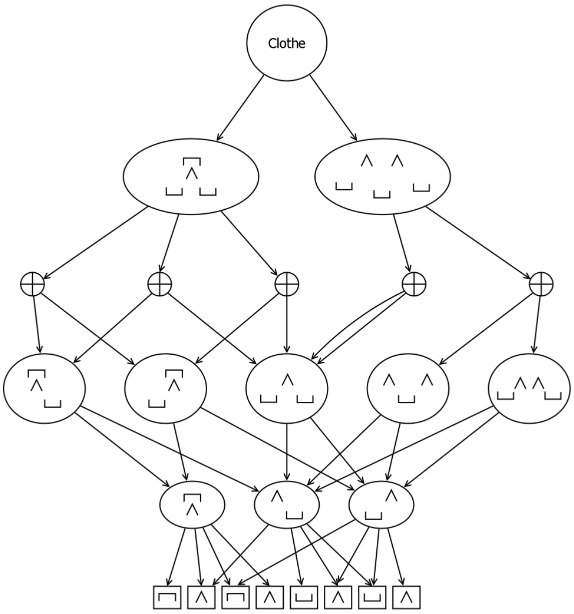
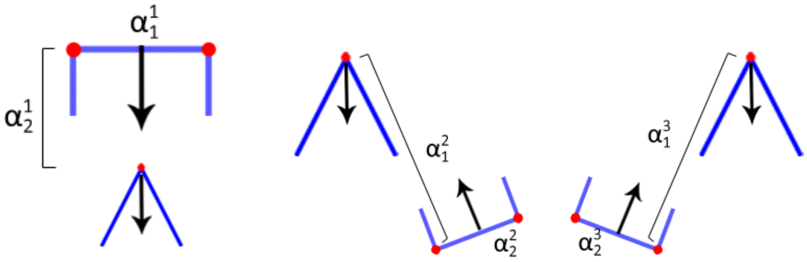


Figure 22: Graph grammar of the implementation. It is used to recognize trousers and shirts.



(a) (b) (c)

Figure 23: a-c) Rules 1-3 along with their attributes.

We denote with $a^i_j$ the $j$th attribute of the $i$th rule or configuration. Table 2 displays all the attributes used so far. They are also shown graphically in Figure 23 for the first three rules and they are identical for all the other rules.

In the *'Trousers'* and *'Shirt'* rules all attributes are inherited from their children. In the next section we describe how we use these attributes in order to infer the types of clothes.

## 6.5 Inference, Training and top-down hypothesis

The inference of the model for a specific image has the goal of validating the bottom-up proposals for clothes types by a top-down strategy. Our current inference strategy is based on heuristic rules, however, it produces promising results on our clothes database. At the moment, our database contains clothes which are mainly unfolded, or have at most one

Table 2: Attributes of the graph nodes

| Attribute | Description |
|---|---|
| $a_1^1$ | The distance between the two corners of the “Π”. |
| $a_2^1$ | The vertical distance from the top of the “Λ” to the horizontal line of “Π”. |
| $a_1^2$ | The distance between the top of the “Λ” and the closest corner of the “Π”. |
| $a_2^2$ | The distance between the two corners of the “Π”. |
| $a_{1,2}^3$ | Same as $a_{1,2}^2$ |
| $a_{1,2}^4$ | Same as $a_{1,2}^1$ |
| $a_{3,4}^4$ | Same as $a_{1,2}^2$ |
| $a_{1,2}^5$ | Same as $a_{1,2}^1$ |
| $a_{3,4}^5$ | Same as $a_{1,2}^3$ |
| $a_{1,2}^6$ | Same as $a_{1,2}^2$ |
| $a_{3,4}^6$ | Same as $a_{1,2}^3$ |
| $a_{1,2}^7$ | Same as $a_{1,2}^2$ |
| $a_3^6$ | Same as $a_2^3$ |

fold. Example images are shown in Figure 24. In order to infer how well a rule applies to a certain garment, we observe some characteristics on the attributes of the rules.

When we have an image of trousers, rule 1 can be applied. In all trousers, the ratio $a_1^1/a_2^1$ takes values inside a small range. So this ratio can be used in order to assess how well *rule 1* corresponds to trousers. Furthermore, rules 4 and 5 can be applied. It is also observed that $a_4^4/a_1^4$ and $a_4^5/a_1^5$ take values around $1/2$. Attributes $a_4^4$ and $a_4^5$ correspond to the width of the leg, while attributes $a_1^4$ and $a_1^5$ correspond to the length of the waist. Also in rule 6, $a_1^6$ and $a_3^6$ corresponds to the length of each leg. We can assess this rule by examining how close the ratio $a_1^6/a_3^6$ is to 1. Finally, in order to discard quickly some instances of the above rules produced by noise, we check the width and height of the legs to take reasonable values for a human body.

Considering the shirts on the other hand, we examined rule 7. As in the previous example, again the ratio $a_1^7/a_2^7$ takes values inside a limited range. Also $a_1^7$ and $a_3^7$, being the two vertical sides of the main body part, should take almost equal values. Moreover, we check again the width of the sleeves and the main body part to take reasonable values according to human body.

Training consists of learning all the above ratios (or any other production rule used) for every type of garment. Having enough values obtained from training, we can use a

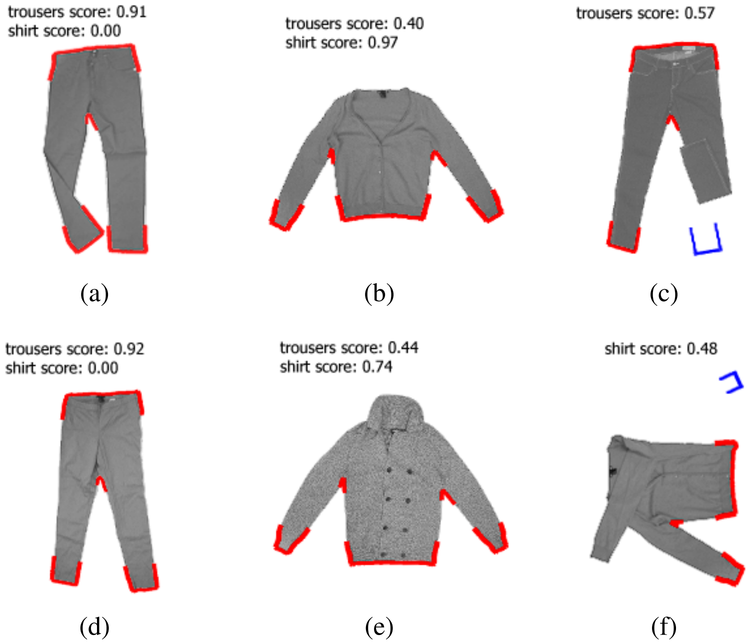


Figure 24: a-c) Rules 1-3. d-f) Rules 1-3 examples on trousers.

simple nearest-neighbour inference algorithm to assess the type of clothes. For training we used half of the images in the database, previously annotated. Images contain only trousers and shirts.

Another important extension of our architecture is the ability to create hypothesis for parts of the clothes that are missing in our graph grammar. If a part is missing among our primitives, we cannot parse the graph until we reach our final rules and we cannot gain a complete clothes configuration. Intermediate nodes which contain sub-configurations of a cloth, can then be used to assess the possible positions of missing primitives. Primitives can be missing for mainly two reasons. The cloth is folded so there are self-occlusions and misplacements or image processing techniques failed to retrieve some primitives. Based on geometric rules, we can make initial hypothesis of the missing parts of the configuration. This can be helpful in order to search again for primitives in a smaller area of interest, with different parameters. On the other hand, we can predict folds and give instructions to the robot in order to manipulate the clothes and bring them to a known configuration.

In Figure 24 some results of clothes recognition are presented. In our database, we obtain 100% success in recognizing simple, mainly unfolded clothes. Moreover, Figures 24(c) and 24(f) show two examples of folded clothes, and how the position of the missing parts can be predicted.

## 6.6 Future Work

Working towards the development of the final graph grammar architecture we plan to gradually increase the complexity of the tree structure presented in Figure 22. First of all we will expand the rules and attributes, so that other clothes types can also be addressed. Of course, this means that the primitives extraction process should be revisited and enriched with interior features, like buttons, collars, pockets, etc. We also plan to investigate the possibility of using the data acquired from the 3D sensors so as to enrich even further the primitives list and the formulation of the logic rules.

An even more promising framework for this architecture would be an automated formulation of logical rules, based on training data.

A large portion of our efforts will also be dedicated to link our architecture with the clothes manipulation process. When ambiguities arise in the graph parsing and no clothes configuration can be inferred, then we can make hypothesis according to the rules. Through robot interaction, we will try to extract any occluding parts of the clothes and manipulate them so that an adequate information may become available for successful recognition.

Finally, it appears that the And-Or graph grammar is sufficiently general to allow incorporation of alternative robot 'action' nodes as a way of actively reasoning about the clothes manipulations and thus obtaining missing information in a top-down direction. This will be investigated as a way of possibly integrating the sensory and the motor parts of the project.

# 7 Conclusions

The deliverable reports our progress on the perception of garment surfaces aimed at their efficient manipulation. Unlike robotic manipulation of rigid objects, the huge configuration space of deformable objects such as clothes, requires more powerful perception. Noticing that humans break down the problem into a set of increasingly simpler tasks (with respect to vision) we follow a similar line of attack.

The first work presented deals with recovering the configuration of a hanging towel from a single image. We have used a simple representation of the hanging cloth as a set of overlaping surfaces together with a connected graph of edges and junctions on the image. We have demonstrated that by using simple heuristics obtained from prior knoweldge it is possible to rule out invalid configuration hypotheses and thus obtain one or two possible configurations. Although this is a step forward with respect to the state-of-the-art there are also several limitations that we wish to investigate in the future. Our approach relies on relatively good segmentation of the surface, and our current feature extraction (fold and corner detection) and topological network construction are not robust enough. We believe that these may be improved using more elaborated region-based segmentation techniques, combining color and depth information, and using less noisy depth images (by means of the CloPeMa stereo-head). Furthermore our approach is currently limited to a rectangular template. However, the geometric rules used may be extended to an arbitrary known polygon (e.g. a shirt template) and even if the template is not known one may iterate over the few possible garment shapes (at least under the assumption that it is topologically planar).

The second work regards recognition of the garment and registration with a template. Our first results were on the use of state-of-the-art deformable shape matching techniques for matching a garment image in a known configuration with a set of templates. We obtained very good results on a realistic data set. Later on we investigated the challenging and novel problem of recognizing the shape of a garment with unknown configuration but limited to a one or more planar folds. Our preliminary results indicate that partial matching techniques combined with geometric inference were very effective for recognizing the shape of the garment and subsequently recovering the location of the fold(s). We are currently testing this approach on a data set of real clothe images. Although the performance is expected to drop there is still plenty of room for improvement since for the moment only boundary information has been exploited.

The third research strand is on understanding the high-level structure of garments. A graph-grammar approach was proposed that helps to combine high-level “syntactic” knowledge with low-level features. Our goal is two-fold. First, we aim at recognizing garments even when only a few cues are visible on the image (e.g. humans will recognize a shirt only be seeing a collar and a few buttons). Our framework allows for incrementally incorporating such cues (also tactile, movement) thus may be applied at all stages of the manipulation. Secondly, in order to perform tasks such as folding that are defined by high-level actions, we need to obtain a high-level representation (e.g. sleeves, button-line, colar). For the moment we have demonstrated the feasibility of our approach on relatively simple examples. However, there are several limitation that still have to be addressed. The most important one is the need for enriching the compendium of visual cues that are currently used so that we may exploit the presence of structures on the surface of the garment (e.g. pockets, necklines). We are working on a machine-learning based approach that will allow us to incorporate such cues in the framework without resorting to heuristics.

Finally, let us bring the above in the context of the laundry folding task (in a simplified manner). The clothe is initially grasped by a corner. The perception system will produce a list of possible configurations and associated grasp affordances. Alternatively if no grasp affoardances are identified a rotation or re-grasping strategy will be proposed. Once a grasp affordance has been identified it will be grasped with the other hand and the clothe will hopefully become flat. Then the clothe will be laid on the table, and the shape registration/recognition algorithm may be applied. An unfolding strategy may now be planned and performed. The high-level perception module will infer garment parts and the folding routine will be initiated.

# References

[1] P. F. Alcantarilla and A. Bartoli. Deformable 3d reconstruction with an object database. In *BMVC'12'*, 2012.

[2] D. Baraff and A. Witkin. Large Steps in Cloth Simulation. In *SIGGRAPH '98*, pages 43–54, 1998.

[3] A. Bartoli, Y. Gerard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[5] Christian Bersch and Benjamin Pitzer. Robotic Cloth Manipulation using Learned Grasp Evaluations. *Workshop on Mobile Manipulation*.

[6] Kiran S. Bhat, Christopher D. Twigg, Jessica K. Hodgins, Pradeep K. Khosla, Zoran Popović, and Steven M. Seitz. Estimating cloth simulation parameters from video. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, SCA '03, pages 37–51, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[7] F. Brunet, R. Hartley, A. Bartoli, N. Nassir, and R. Malgouyres. Monocular template-based reconstruction of smooth and inextensible surfaces. In *Proceedings of the Tenth Asian Conference on Computer Vision (ACCV 2010)*, volume 6494 of *Lecture Notes in Computer Science*, pages 52–66, Queenstown (New Zealand), November 2010.

[8] Michel Carignan, Ying Yang, Nadia Magnenat Thalmann, and Daniel Thalmann. Dressing animated synthetic actors with complex deformable clothes. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '92, pages 99–104, New York, NY, USA, 1992. ACM.

[9] Hatem Charfi, Andr Gagalowicz, and Rmi Brun. Determination of fabric viscosity parameters using iterative minimization. In Andr Gagalowicz and Wilfried Philips, editors, *Computer Analysis of Images and Patterns*, volume 3691 of *Lecture Notes in Computer Science*, pages 789–798. Springer Berlin Heidelberg, 2005.

[10] Hong Chen, Zi Jian Xu, Zi Qiang Liu, and Song Chun Zhu. Composite templates for cloth modeling and sketching. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 943–950, Washington, DC, USA, 2006. IEEE Computer Society.

[11] K-J. Choi and H-S. Ko. Research Problems in Clothing Simulation. *Computer-Aided Design*, 37:585–592, 2005.

[12] T. Collins and A. Bartoli. Locally affine and planar deformable surface reconstruction from video. In *VMV'10 - Proceedings of the International Workshop on Vision, Modeling and Visualization*, November 2010.

[13] T. Collins, A. Bartoli, and R. Fisher. Automatic quasi-isometric surface recovery and registration from 4d range data. In *BMVA Symposium on 3D Video Analysis, Display and Applications*, 2008.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2001.

[15] Marco Cusumano-towner, Arjun Singh, Stephen Miller, James F O Brien, and Pieter Abbeel. Bringing Clothing into Desired Configurations with Limited Perception. *ICRA*, 2011.

[16] M. Desbrun, P. Schrder, and A. Barr. Interactive animation of structured deformable objects. In *SIGGRAPH '99*, 1999.

[17] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, January 2005.

[18] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *in Proc. CVPR*, 2003.

[19] R. Ferreiram, J. Xavier, and J. Costeira. Shape from motion of nonrigid objects: The case of isometrically deformable flat surfaces. In *in Proc. BMVC*, 2009.

[20] N. A. Gumerov, Z. Ali, R. Duraiswami, and L. S. Davis. Structure of applicable surfaces from single views. In *In ECCV04.*, pages 482–496, 2004.

[21] Kyoko Hamajima and Masayoshi Kakikura. Planning strategy for task of unfolding clothes. *Robotics and Autonomous Systems*, 32(August 1999):145–152, 2000.

[22] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing with attribute grammar. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):59 –73, jan. 2009.

[23] N. Hasler, B. Rosenhahn, M. Asbach, J. r. Ohm, and H. p. Seidel. An analysis-by-synthesis approach to tracking of textiles. In *In Proc. of the International Workshop on Motion and Video Computing*, 2007.

[24] S. Hata, T. Hiroyasu, J. Hayashi, H. Hojoh, and T. Hamada. Robot system for cloth handling. In *Industrial Electronics, 2008. IECON 2008. 34th Annual Conference of IEEE*, pages 3449 –3454, nov. 2008.

[25] John Hershberger and Jack Snoeyink. An o(nlogn) implementation of the douglas-peucker algorithm for line simplification. In *Proceedings of the tenth annual symposium on Computational geometry*, SCG '94, pages 383–384, New York, NY, USA, 1994. ACM.

[26] B. Horowitz and A. Pentland. Recovery of non-rigid motion and structure. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 325 –330, jun 1991.

[27] Nebojsa Jojic and Thomas S. Huang. On analysis of cloth drape range data. In *In Proceedings of Asian Conference on Computer Vision*, pages 463–470. Springer-Verlag, 1998.

[28] Manabu Kaneko and Masayoshi Kakikura. Planning Strategy for Putting away Laundry - Isolating and Unfolding Task -. *Symposium on Assembly and Task Planning*, pages 429–434, 2001.

[29] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1988.

[30] L. Kaufmann and P. J. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1 Norm*, pages 405–416, 1987.

[31] Yasuyo Kita, Fumio Kanehiro, Toshio Ueshiba, and Nobuyuki Kita. Clothes handling based on recognition by strategic observation. In *Humanoids*, pages 53–58. IEEE, 2011.

[32] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. A method for handling a specific part of clothing by dual arms. *International Conference on Intelligent Robots and Systems*, 1:4180–4185, 2009.

[33] Yasuyo Kita, Toshio Ueshiba, Ee Sian Neo, and Nobuyuki Kita. Clothes state recognition using 3d observed data. In *ICRA*, pages 1220–1225. IEEE, 2009.

[34] Liang Lin, Xiaobai Liu, Shaowu Peng, Hongyang Chao, Yongtian Wang, and Bo Jiang. Object categorization with sketch representation and generalized samples. *Pattern Recognition*, 45(10):3648–3660, 2012.

[35] Liang Lin, Tianfu Wu, Jake Porway, and Zijian Xu. A stochastic graph grammar for compositional object representation and recognition. *Pattern Recogn.*, 42(7):1297–1307, July 2009.

[36] Haibin Ling and David W. Jacobs. Shape classification using the inner-distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):286–299, 2007.

[37] 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. *Robotics and Automation*, pages 2308–2315, 2010.

[38] D. Metaxas and D. Terzopoulos. Dynamic deformation of solid primitives with constraints. In *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, SIGGRAPH ’92, pages 309–312, New York, NY, USA, 1992. ACM.

[39] Stephen Miller, Jur Van Den Berg, Mario Fritz, Trevor Darrell, Ken Goldberg, and Pieter Abbeel. a geometric approach to Robotic Laundry Folding. *The International Journal of Robotics Research*, 2012.

[40] Stephen Miller, Mario Fritz, Trevor Darrell, and Pieter Abbeel. Parametrized shape models for clothing. In *ICRA*, pages 4861–4868. IEEE, 2011.

[41] Vlad I. Morariu, David Harwood, and Larry S. Davis. Tracking people’s hands and feet using mixed network and/or search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1, 2012.

[42] Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *Proceedings of 4th International Joint Conference on Pattern Recognition*, pages 752–754, 1978.

[43] Fumiaki Osawa, Hiroaki Seki, and Yoshitsugu Kamiya. Clothes Folding Task by Tool-Using Robot. *Journal of Robotics and Mechatronics*, 18(5), 2006.

[44] Fumiaki Osawa, Hiroaki Seki, and Yoshitsugu Kamiya. Unfolding of Massive Laundry and Classification Types. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, pages 457–463, 2006.

[45] D. Pritchard and W. Heidrich. Cloth motion capture. In *in Proc. Eurographics*, 2003.

[46] Arnau Ramisa, Guillem Aleny, Francesc Moreno-noguer, and Carme Torras. Using Depth and Appearance Features for Informed Robot Grasping of Highly Wrinkled Clothes. *ICRA*, pages 1703–1708, 2012.

[47] Hayko Riemenschneider, Michael Donoser, and Horst Bischof. Using partial edge contour matches for efficient object category localization. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 29–42, Berlin, Heidelberg, 2010. Springer-Verlag.

[48] Brandon Rothrock and Song-Chun Zhu. Human parsing using stochastic and-or grammars and rich appearances. In *International Workshop on Stochastic Image Grammar*, 2011.

[49] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

[50] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[51] M. Salzmann, F. Morenoguer, V. Lepetit, and P. Fua. P.: Closed-form solution to non-rigid 3d surface registration, 2008.

[52] Volker Scholz, Timo Stich, Marcus Magnor, Michael Keckeisen, and Markus Wacker. Garment motion capture using color-coded patterns. In *ACM SIGGRAPH 2005 Sketches*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[53] Zhangzhang Si, Mingtao Pei, Benjamin Yao, and Song-Chun Zhu. Unsupervised learning of event and-or grammar and semantics from video. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 41–48, Washington, DC, USA, 2011. IEEE Computer Society.

[54] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[55] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischert. Elastically Deformable Models. *Computer Graphics*, 21(4):457–463, July 1987.

[56] Dimitra Triantafyllou and Nikos A Aspragathos. A Vision System for the Unfolding of Highly Non-rigid Objects on a Table by One Manipulator. *ICIRA*, pages 509–519, 2011.

[57] Chin-Hsien Tseng, Shao-Shin Hung, Jyh-Jong Tsay, and Derchian Tsaih. An efficient garment visual search based on shape context. *W. Trans. on Comp.*, 8(7):1195–1204, July 2009.

[58] Igor Guskov University and Igor Guskov. Direct pattern tracking on flexible geometry. In *In Proc. WSCG*, 2002.

[59] Bryan Willimon, Stan Birchfield, and Ian Walker. Classification of Clothing using Interactive Perception. *Inernational Conference on Robotics and Automation*, pages 1862–1868, 2011.

[60] Bryan Willimon, Stan Birchfield, and Ian D. Walker. Model for unfolding laundry using interactive perception. In *IROS*, pages 4871–4876. IEEE, 2011.

[61] Tianfu Wu and Song-Chun Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International Journal of Computer Vision*, 93:226–252, 2011.

[62] Long Zhu, Yuanhao Chen, Yifei Lu, Chenxi Lin, and A. Yuille. Max margin and/or graph learning for parsing the human body. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[63] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2(4):259–362, January 2006.