Pose and Category Recognition of Highly Deformable Objects Using Deep Learning

Ioannis Mariolis, Georgia Peleka, Andreas Kargakos, and Sotiris Malassiotis
Information Technologies Institute, Centre for Research & Technology Hellas
6th km Xarilaou-Thermi, 57001, Thessaloniki, Greece
Email:{ymariolis,gepe,akargakos,malasiot}@iti.gr

***Abstract*—Category and pose recognition of highly deformable objects is considered a challenging problem in computer vision and robotics. In this study, we investigate recognition and pose estimation of garments hanging from a single point, using a hierarchy of deep convolutional neural networks. The adopted framework contains two layers. The deep convolutional network of the first layer is used for classifying the garment to one of the predefined categories, whereas in the second layer a category specific deep convolutional network performs pose estimation. The method has been evaluated using both synthetic and real datasets of depth images and an actual robotic platform. Experiments demonstrate that the task at hand may be performed with sufficient accuracy, to allow application in several practical scenarios.**

I. Introduction

Autonomous manipulation of highly deformable objects, such as garments, is a very challenging task in the domain of robotics. In order to enable robots to manipulate such objects, recognition of their deformed state should be performed using robust computer vision algorithms. Although some results have been already demonstrated when deformation is rather small [1], [2], [3], interpreting highly deformed states remains elusive. In this work, the deformation state space is reduced by picking and hanging the garment from a random point. But even in this case, the investigated state space remains huge, making impractical the use of shape matching techniques and typical machine learning approaches. An additional challenge is introduced by the large variety of shapes, materials, sizes and textures of real garments. Luckily, with the use of a 3D sensor, we may disregard variations caused by texture or illumination.

In this paper, we are addressing the problem of recognizing the category and pose of hung garments, as they are grasped by a robotic manipulator (Fig. 1). The output of our method can be used in many scenarios, such as autonomous sorting of garments for laundry or recycling, as well as to aid manipulation, eg. unfolding, laying the garment flat on a table, etc. These scenarios are currently widely applicable in industry, but they may soon be performed at home by future domestic robots. We propose the use of a hierarchy of deep convolutional networks trained by depth images of hung garments to first recognize the garment's category and then estimate its current pose. In that direction, we use a low-cost depth sensor to acquire a large dataset of exemplars. In addition, we employ physics based simulation of hanging garments to acquire an even larger synthetic dataset. The acquired depth images are used as raw input to deep Convolutional Neural Networks (CNNs), that learn discriminant representations, suitable for recognition and pose estimation. CNNs are widely used in image recognition tasks [4], [5], [6], [7], whereas recent success in outperforming state of the art at benchmarks such as the MNIST database and ImageNet [8] using deep architectures of CNNs, has increased their popularity. In very recent studies [9], such networks have been proposed for the related problem of human pose recognition, presenting very promising results.

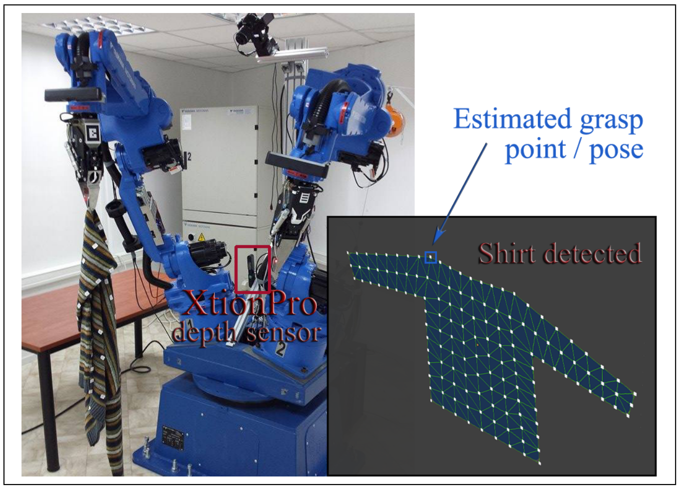

Fig. 1. Investigated application scenario: a Motoman robot grasps a garment, and an XtionPro sensor acquires depth images to recognize the garment's type and pose.

Other studies that address the same problem with us are [10], [11], [12]. They are all based on the use of simulated data of hung garments. The most similar work to ours is [11], where SIFT-based features are extracted by synthetic depth images and an hierarchy of Support Vector Machines (SVMs) is trained to recognize the category and pose of hung garments. In comparison to these works our key contributions are:

- Formulation of a real-time deep learning scheme that, i) extracts discriminant features from raw depth images of hung garments, ii) uses the learned features for recognizing the category and pose of the garments.
- The proposed system is faster and more robust than

978-1-4673-7509-2/15/$31.00 2015 IEEE

the one in [11], that uses hand-engineered features and SVMs. It is also more accurate in pose estimation than both [11] and [12]. As opposed to these works, acceptable performance is reported by our system even without aggregating single-view classification results.

- Investigation of the use of synthetic data for designing and pre-training large and deep networks that perform well even with smaller real datasets. In our simulations we have used simple 2D models of garments, whereas in [11] and [12] commercial 3D models very similar to the real garments used for testing are employed.
- Extensive experimental results on large datasets, demonstrating beyond state of the art performance.

II. Related Work

It can be argued that recognition and manipulation of deformable objects has been rather under-explored by the computer vision and robotics community, with initial works employing template matching techniques, ad hoc rules and heuristics [13], [14], [15]. The field is drawing more attention, especially after the recent success of a number of works [16], [17], [18] on garment recognition and manipulation, using a PR2 robot on tasks such as clothes folding. However, the employed recognition methods are mostly based on aligning an observed shape/contour to an existing one. An active vision approach is employed by Willimon et al. in [19] and [20] for hung garment recognition, and recognition of crumpled garments lying on a table, respectively. Limited interaction with the garment is performed, such as dropping and regrasping for the hanging case, and pulling in predefined directions for the clumped case, and various visual features are extracted. However, the employed feature classification methods heavily depend on color-based segmentation, limiting their applicability in case of garments with high texture variance. Doumanoglou et al. [21], [22] also employ an active vision approach in order to unfold a crumpled garment grasped by its lowest hanging point. Before regrasping the hung garment, its category is recognized using Random Forests. Training is performed using small feature vectors extracted by random pixel tests on depth images acquired using an XtionPro sensor. High recognition rates are reported, but the method relies on a lowest hanging point heuristic. Although some of the above studies infer the garments' configuration at some point during manipulation, they don't estimate the pose of the garment when it is hung by a single arbitrary point. This problem is explicitly addressed by Kita et al. [23], [24] in a series of works, using stereo vision to extract 3D information about the hung garment's configuration. In a more recent work, Kita et al. [10], a trinocular stereo system is employed for extracting a 3D model of the garment, and its state is recognized by comparing the model with candidate shapes, which are predicted in advance via simulation. The works [11], [12] by Li et al. are also based on the use of simulated models of hung garments in order to recognize their state. However, instead of relying on optical sensors, they employ a Kinect depth sensor to acquire 3D information of the garment's configuration. More specifically, in [11] a hierarchical classifier using two-layers of SVMs is used for first recognizing the category, and then estimating the location of the grasping point on the garment (pose). The SVMs are trained using SIFT features extracted by the depth images of simulated garments and compressed using sparse coding. Simulated garments are also employed in [12], but instead of using the simulation results to acquire synthetic depth images, the opposite is proposed. Namely, the depth images of the real garments are used for constructing a 3D model that is then matched to the most similar synthetic model. Matching is based on a weighted Hamming distance metric, which is learned using a small number of labelled poses of the real garments. The last two studies are the most similar to our work and their results are compared with the ones produced by our method.

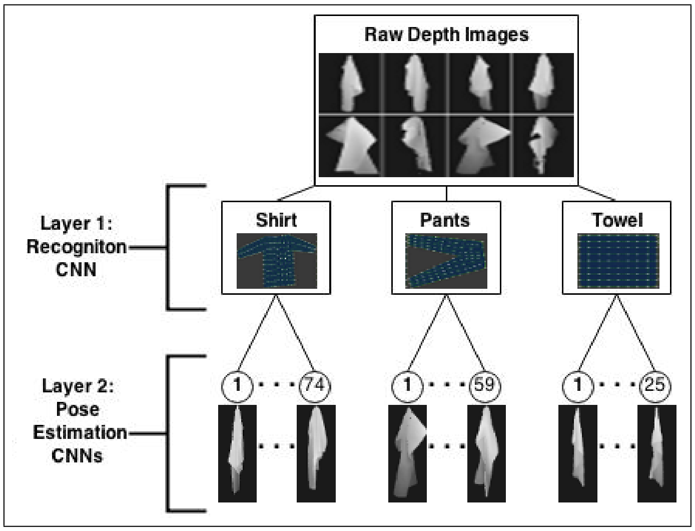

Fig. 2. Overview of the proposed hierarchical approach. In the first layer the hung garment is recognized by a deep CNN. Then, a category specific CNN is performing pose estimation.

In this paper, we are using depth images of the hung garments as input for deep CNNs. Adopting the hierarchical approach used in [11], our classifier contains two layers. In the first layer, a deep CNN is used for classifying the hung garment to one of the examined categories, which in our case are shirts, pants, and towels. In the second layer, another deep CNN with similar architecture is employed for inferring the pose of the hung garment by estimating the grasping point location on a garment template. In the second layer, different CNNs are used for inferring the pose of garments belonging to different categories. Hence, the output of the first layer is used for selecting the appropriate CNN for the second layer. An overview of the proposed method is presented in Figure 2.

III. Materials and Methods

Images of hung garments are expected to exhibit stationarity of statistics and locality of pixel dependencies. These properties are more evident in case of depth images, since texture

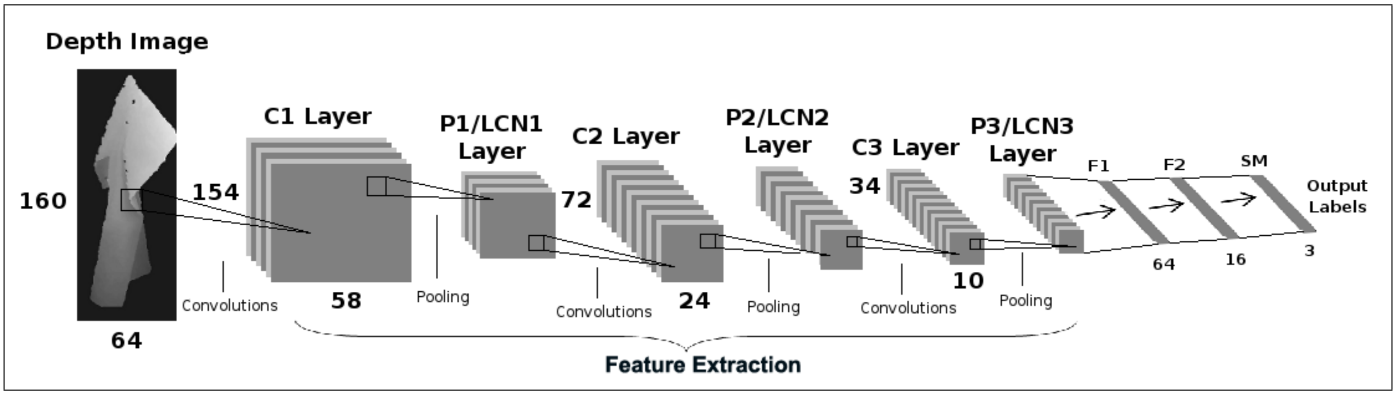

Fig. 3. Network architecture for recognition CNN. The network consists of 3 Convolutional layers and accompanying Pooling layers, followed by 2 Fully-connected hidden layers. A softmax output layer is added after the last Fully-connected layer.

information is discarded and local dependencies attributed to the garment's configuration are enhanced. Therefore, in order to perform category and pose recognition of hung garments, it makes sense to employ a learning scheme where the depth images are used as inputs to CNNs, since their success builds strongly on these assumptions [8]. However, the extremely large state space of hung configurations and the immense variety of garment types, sizes, and material properties call for deep CNN architectures, where many convolutional layers are successively employed to automatically learn discriminative representations of the garments. The above challenges also motivate the adoption of a hierarchical approach, where the category of the garment is first identified and then a different CNN is employed for inferring its pose. In this work, three garment categories are considered: shirt, pants, and towels. Hence, the second layer of our hierarchical classifier is consisted by three different CNNs. Since different number of poses is considered for each category, the main difference between the 2nd-layer CNNs is the number of neurons in their output layer.

A. Depth Data Acquisition

In order to acquire the depth input images, an ASUS XtionPro depth sensor, which is similar to Kinect, is employed. The depth sensor is mounted on the robot's base, and the hung garment is positioned by the robot in front of the sensor, at a distance of approximately 1 m (see Figure 1). Then, the robot gripper starts rotating around the vertical axis and a series of depth images are successively acquired, capturing overlapping views of the rotated garment until a full rotation is performed. At that point, about 180 images have been captured and the gripper retrieves its original position. In order to minimize oscillations, rotation speed is kept low, increasing the time needed for capturing the images for each pose, slowing acquisition down. Therefore, apart from real world data, synthetic 2D models of garments have been constructed (see Figure 5) and a significantly greater amount of depth images has been acquired, using virtual scenes that replicated to a large extend the real setup. The main motivation behind the creation of the synthetic dataset is to use it for guiding the design of the CNNs architectures. Since, simulation can provide a plethora of training data, the evaluation of larger and deeper networks and their comparison to simpler architectures becomes feasible. At the same time, cross-domain knowledge transfer can be investigated using a combination of synthetic and real datasets for training the deep networks.

B. Deep Learning Models

Using the synthetic dataset several CNN architectures were investigated (see next Section for details), and the selected ones for category recognition and pose estimation are presented here.

Deep CNN for Category Recognition The network's architecture is depicted in Figure 3. It contains five layers with learnable parameters, where the first three layers are convolutional and the remaining two are fully-connected. The output of the last fully-connected layer is connected to a 3-way softmax which estimates a probability distribution over the 3 garment categories. Let C denote a convolutional layer, P an L2-norm pooling layer of stride 2, LCN a local contrast normalization layer [4], and F a fully-connected layer. Both C and F layers consist of a linear transformation followed by a nonlinear one, which in our case is the hyperbolic tangent non-linearity. For C layers, the size is defined as $width \times height \times filters$, where $width$ and $height$ have a spatial interpretation, whereas $filters$ denotes the number of applied convolutional kernels. If we use parentheses to include the size of each layer, then the network description can be given as $C(154 \times 58 \times 16)$ _ P _ LCN _ $C(72 \times 24 \times 32)$ _ P _ LCN _ $C(34 \times 10 \times 16)$ _ P _ LCN _ $F(64)$ _ $F(16)$ _ $SM(3)$, where SM denotes the softmax output layer. The filter size for the first C layer is 7×7, for the second 5×5, and the third 3×3. The input to the net is a depth image of 160×64 pixels. The total number of parameters in the above model is about 112K.

As suggested by the use of the softmax layer, the network is trained (in a supervised fashion) by minimizing a Cross Entropy objective function, whereas the minimization

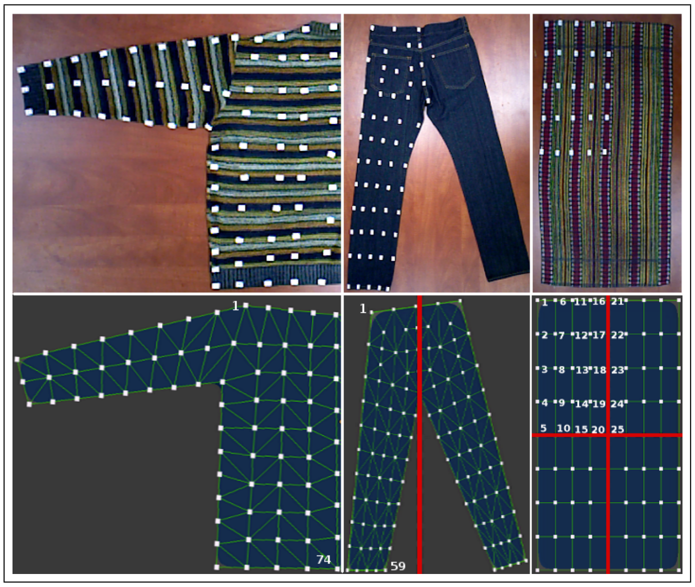

Fig. 4. Selected vertices defining the poses of the hanging garments. In the top row, white markers define the poses on real garments. In the bottom row, the synthetic models are depicted. The selected poses are defined by their vertices that reside to the left of their symmetry axes.

is performed by means of Stochastic Gradient Descent (SGD) optimization. The network's performance evaluation is based on reported Correct Classification Rate (CR).

Deep CNN for Pose Estimation In this study, similarly to [11], pose estimation is performed by means of classification, with each pose defined as a grasping point on the garment's surface. Thus, a similar architecture to recognition CNNs is also adopted for the CNNs performing pose estimation. However, in this case a larger number of classes (poses) is defined for each garment type. In case of shirts there are 74 poses, in case of pants 59, and in case of towels 25. These poses have been defined to coincide with the vertices of the simulated models after discarding symmetric counterparts, and are denoted, by Arabic numbers on the models of Figure 4. In this figure, only the towel labels are exhaustively displayed. Notice that towels present two symmetry axes, resulting to a smaller number of different poses.

The networks used for pose estimation can be described as $C(154 \times 58 \times 32)$ _ P _ LCN _ $C(72 \times 24 \times 64)$ _ P _ LCN _ $C(34 \times 10 \times 32)$ _ P _ LCN _ $F(256)$ _ $F(128)$ _ $SM(n)$, where n is the number of poses and depends on the garment category. Hence, the topology and the depth of the networks is identical to the recognition CNN, but the size of the network increases to compensate for the large increase in the number of output classes. The total number of parameters in the above model is over 832K, i.e. 8 times higher than the recognition network. Although training is performed in the same fashion as in recognition, evaluation of the network's performance is not based, in this case, on the reported CR. Since we expect the garment to present similar configurations when grasped by neighbouring points, it makes more sense to use metrics suitable for regression. A natural choice could be the Geodesic distance between the predicted and the actual pose (grasping point). However, since the poses have been defined on the 2D synthetic models an acceptable approximation is to use the Inner-distance metric [25]. Given a shape O and two points $x, y \in O$, the inner-distance between x, y denoted as $d(x, y; O)$, is defined as the length of the shortest path connecting x and y within O. O is defined as a connected and closed subset of R^2. Thus, the Inner-distance between predicted and the actual pose, called henceforth Error Distance, is computed, and its mean over the test set is used for assessing the performance of the network. Adopting the approach of [11] and [12], the distribution of the Error Distance is also estimated using the percentage of the tested grasping points that reside within that distance.

C. Aggregation of Viewpoint Classification Results

The presented CNNs perform single-view classification for both category and pose recognition. However, the acquisition setup allows aggregation of the single-view results for the entire dataset of the 180 depth images that correspond to the same pose. A simple but effective method to achieve this is to perform majority voting between the different outputs of the 180 single-view classifications. This approach can be applied to both category and pose classifiers, boosting the performance and introducing more robustness to the classifiers.

IV. Experimental Evaluation

A. Setup

Robotic manipulations have been conducted using a dual arm robot composed by two M1400 Yaskawa arms mounted on a rotating base. The depth images of the hanging garments have been acquired by an Asus XtionPro depth sensor placed between the arms at a fixed height. Grasping has been based on custom made grippers [26].

Synthetic Dataset Before evaluating our method on real data acquired using the above setup, we have constructed a large dataset (SD) of synthetic depth images using Blender 2.6.2, an open-source 3D computer graphics software. We have constructed 24 models of shirts, 24 models of pants, and 24 models of towels. The models of the same category differ in shape, size and material properties. In order to simplify and speed-up the simulation process we are using, as in [10], 2-D models of the garments, assuming that the front and back sides of the clothes are not separated. Even with this simplification, the models approximate surprisingly well the configuration of real garments hung under gravity (see Figure 5). An example model for each category along with their corresponding triangular mesh, is presented in Figure 4. The mesh of the shirt models consists of 141 vertices, whereas there are 113 vertices for pants and 81 vertices for towels. Due to symmetry the above vertices correspond to 74, 59 and 25 distinct poses respectively. Using Blender's cloth engine, we have simulated hanging the models by each of their vertices. A ring of 80 virtual cameras is placed uniformly around the hung garment in such height that they can view the entire garment, and 80 synthetic depth images are acquired. The use of the camera ring depletes the need for rotating the garment, which introduces additional noise during acquisition with the real system that uses only a single sensor. The formulation of the synthetic dataset needs only a small fraction of the time needed by the real world system, whereas the resulted images are noise-free. The set of the 80 images for a single vertex is acquired in about 1 min. Namely, the entire database of the 643200 depth images has been created in less than a week. As explained in Section 2, the synthetic dataset allows the evaluation of larger and deeper networks and is used for guiding the design of the CNNs architectures. At the same time, it is used for investigating cross-domain knowledge transfer. To facilitate the above tasks, the setup of the virtual cameras should match as close as possible the robotic platform's setup. Thus, the virtual cameras where positioned at a distance of about 1060 mm from the hanging axis, whereas their vertical distance from the hanging point was set to 400 mm equal to the one of the robot's XtionPro sensor. The virtual camera's intrinsic parameters where also matched to those of the XtionPro sensor. Figure 5 presents synthetic depth images of different models hung by their first pose. The views are matched in order to facilitate the comparison between the resulted configurations of models that belong to the same category. It is clear that the differences in the model parameters reflect to differences to the final hung configurations. Hence, it can be assumed that the synthetic dataset presents adequate intra-class variability even in case of pose estimation.

Real Garments Dataset As opposed to the synthetic dataset the acquisition of real data is time-consuming and noisy. The robot grasps a crumbled garment lying on a table and hangs it in front of the XtionPro sensor. Then, the hung garment is rotated 360°, while the sensor is acquiring both RGB and depth images. Image size is 640x480 pixels for both modalities, whereas depth images are thresholded keeping only values between 700 and 1500 mm in order to segment the garment from the background. A dataset, called henceforth RD1, consisting of 4757 depth images has been acquired, with the robot autonomously picking up the garment using curvature cues from the crumbled garment's surface. A different XtionPro sensor, mounted on the robot hand, is employed for capturing a depth image that is used by the pick-up algorithm before grasping. In total, 13 real sized garments have been employed, belonging to the three specified categories. Each garment has been hung 4 or 5 times and about 80 depth images have been acquired for each case. In this autonomous setting, the selection of the pose is random and the acquired dataset is useful for the recognition task only. In order to acquire data for the pose estimation task, a more controlled setting has been selected. Before grasping, markers have been placed on the garments approximately matching the topology of the pose vertices in the simulation. Then, robot grasping is performed manually, with the robot holding the garment from each marker. Acquisition is performed in the same fashion as in the previous case, with the difference that 180 depth images are acquired per pose. In any case, during garment rotation the angular speed of the gripper is held low, in order to reduce oscillation effects. Thus, an average time of about 5 min is necessary to capture the 180 depth images of each pose. For the new dataset, called henceforth RD2, 9 real sized garments have been employed, three for each category, yielding a total of 85320 depth images. An example of the resulted configurations for RD2 and the first pose of each category is presented in Figure 5. The complete database containing both real and synthetic labeled datasets is publicly available[1].

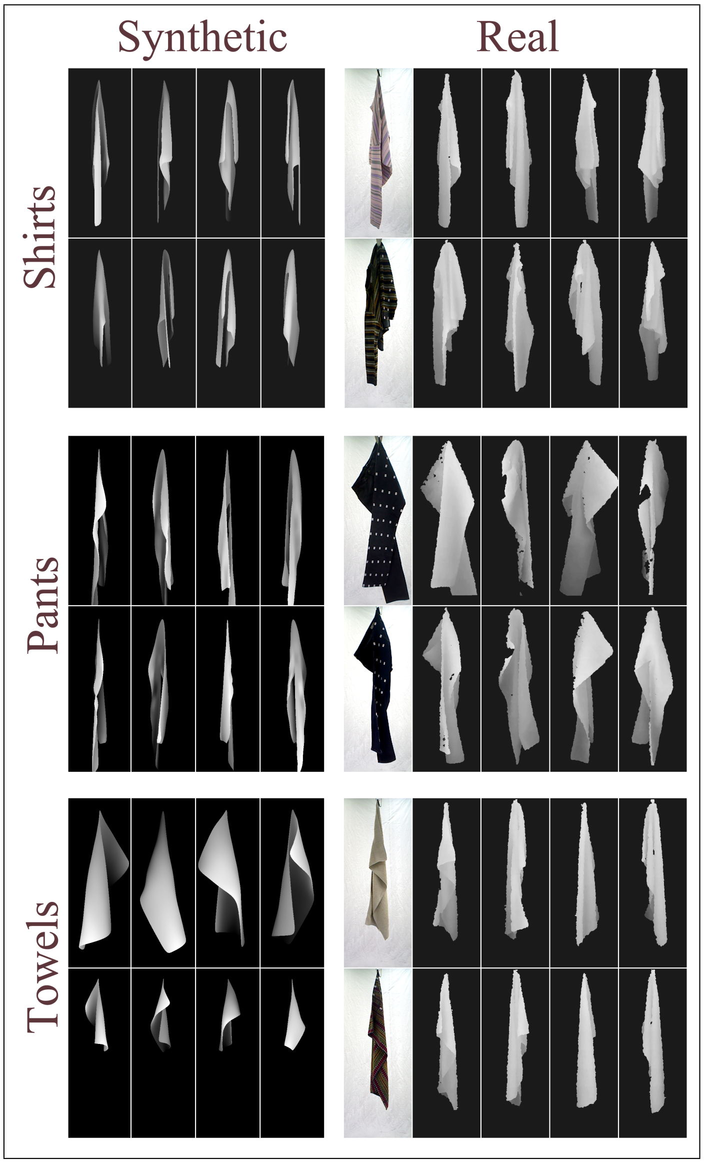

Fig. 5. Example images of synthetic (left) and real (right) garments hung by their first pose.

Experimental Details The implementation and evaluation of the proposed method has been made in Torch7 [27], which is a machine learning library that extends Lua. It provides

[1] http://clopema.iti.gr/datasets/DeepGarmentRecognition

Type	Training	Test	Validation
Shirt	135360	90240	45120
Pants	108480	72320	36160
Towel	77760	51840	25920

TABLE I
DATASET SIZES FOR POSE ESTIMATION USING SD

a flexible environment to design and train learning machines such as the CNNs used in this study. Since in the case of the synthetic dataset we have sufficient amount of labelled data, we have employed a hold out approach and split SD to a training, a test, and a validation set. For the recognition task the training set uses 36 of the synthetic models, the test set 24 models, and the validation set 12 models, yielding 160800 (keeping 40 out of 80 depth images for each pose), 214400, and 107200 depth images, respectively. For the pose estimation task, different training sets are used for each garment type. Each training set uses 12 of the synthetic models, each test set 8 models, and each validation set 4 models. In Table I the corresponding sizes of the datasets used for pose estimation are presented, separately for each garment category. We have used the validation sets for determining optimum learning parameters. The networks performed well for a large range of parameter values. Thus, selecting values within this range, the networks' learning rate was set to 0.001, whereas momentum was set to zero. SGD was performed using batches of a single training instance and early stop at 10 epochs has been applied. For the real dataset, due to their smaller size K-fold cross-validation [28] is employed in a stratified fashion, where all instances of a garment are in the same fold. For RD1 K was set to 10, whereas for the larger RD2 K was set to 3 folds.

Images with initial size 640 × 480 were automatically cropped to 640 × 256 with the garment remaining at the center of the image. Cropping did not result to any loss of information, since when hung, the length of the garment is a lot greater than the width due to the effect of gravity. The cropped images were finally down-sampled to 160 × 64. Training of a single image takes about 12ms in a system with Intel Core i7-4770K CPU @ 3.50GHz 8 processor and 32 Gb RAM, whereas testing takes about 5ms. Thus, even for the viewpoint aggregation approach, the hierarchical classifier is very fast, needing about 1.8 seconds to test the acquired images and produce the category and pose results.

Selecting Network Architecture We began investigating various CNN architectures for category and pose recognition using the validation set of SD. Initially we have trained shallow architectures with only two convolutional and pooling layers and one hidden fully connected layer. Then, we started gradually increasing the number of layers until no significant improvement is presented in the classification accuracy. The architectures described in Section 3 were finally selected, presenting at least 0.5% higher accuracy than the simpler architectures and not less than 0.5% than the deeper architectures that have been examined. The deepest architecture that has been considered employed 4 convolutional layers, 4 pooling layers and 2 fully connected hidden layers.

In accordance to published literature, for the majority of the examined architectures CNNs trained fast. Namely, most CNNs converged to their maximum accuracy in less than 10 epochs, when Stochastic Gradient Descent (SGD) algorithm was used for optimization. The CNNs also presented robustness with respect to the selected learning rate, which for most networks could span several orders of magnitude without a noticeable impact on the performance.

B. Results and Discussion

Category Recognition Using the synthetic dataset for training the CNN, a recognition rate of 92.31% has been reported for the synthetic test set. Then, we retrained the CNN from the beginning using RD1 and 10-fold cross-validation. In that case, an average (over the ten folds) recognition rate of 89.38% is reported. Thus, even though 37 times less data were used the networks' performance was decreased by only 3%. What is more, in case a majority voting scheme is employed for aggregating the results of all the views of each garment in the same pose, the recognition rate for RD1 increases to 94.83%, which can be considered acceptable for our current application.

In order to test whether cross-domain learning can be achieved, we have also used RD1 to test the CNN that was trained by the synthetic data. However, the reported recognition rate was not above chance, implying that some fine tuning of the network is needed before testing with the real data. Since we are using SGD, training is performed on-line, allowing pre-training the network with the synthetic data and continue training with RD1 data. In that case, although both training and testing recognition rates drastically increase in the first epochs (about 20% - 30% in each fold), in the last epochs testing recognition rate reaches 89.51%, which is only slightly higher to the one without pre-training. However, when the results are aggregated over the views a rate of 96.55% is achieved, which is about 2% higher compared to the aggregated results without pre-training. This implies that pre-training helps the network to produce more coherent results for different views of the garment in the same pose.

In order to establish a baseline, the single view performance of our CNN trained with RD1 has been compared to other popular learning schemes such as Support Vector Machine (SVM) and Random Forests (RF), which, however, need as input more discriminant features than the raw depth-images. For that purpose we have extracted Histograms of Oriented Gradients (HOGs) from the depth images and performed learning. The length of the extracted feature vectors was 4788 and the results are presented in Table II. We used one versus one SVM with linear kernels, whereas RFs were trained using at most 100 trees and maximum depth 25. In the last line of the table, PT-RF denotes the learning scheme proposed in Doumanoglou et al. [21]. This scheme is also based on Random Forests, but instead of HOGs the features are extracted by applying certain Pixel Tests (PTs) on the depth image. In that work the method was applied for hung garment recognition and regrasping point estimation. Although related

Learning scheme	Recognition Rate %
CNN	89.38
HOG-SVM	86.41
HOG-RF	83.83
PT-RF [21]	82.05

TABLE II
COMPARATIVE RESULTS BETWEEN PROPOSED METHOD (CNN) AND OTHER WIDELY USED LEARNING SCHEMES SUCH AS SVMS AND RANDOM FORESTS.

Type	$Li14a$	$Li14b$	CNN_{SD}	CNN_{RD2}
Shirt	16.05	13.61	12.62 (13.4)	7.84 (10.95)
Pants	10.89	9.70	6.03 (7.53)	4.61 (7.42)
Towel	N/A	N/A	11.62 (11.81)	1.96 (3.34)

TABLE III
COMPARISON ON MEAN ERROR DISTANCE FOR DIFFERENT TYPES OF GARMENTS. THE EMPLOYED UNITS ARE CENTIMETRES, AND VIEWPOINT AGGREGATION IS APPLIED. IN THE PARENTHESES, SINGLE VIEW RESULTS ARE REPORTED FOR OUR METHOD.

to our work, in that study the pose of the hung garment is restricted to be one of the lowest hanging points. Thus, we do not compare with the results of that study. Instead, Table II contains only the results of applying PT-RF on our dataset. According to these results, our approach outperforms those based on hand engineered features by at least 3%.

Since the convolutional layers of our deep network are learning discriminant representations of the raw depth inputs, it makes sense to also test the performance of these automatically extracted features on a different classifier. Thus, we have taken the output of the last pooling layer of the CNN and used it as a feature vector for a linear SVM. This resulted to a recognition rate of 88.66%, outperforming HOG-SVM by more than 2%, verifying the discriminating power of our learned features.

In case RD2 is employed for training and testing the recognition CNN, a 93% average recognition rate is achieved for single views. Thus, RD2 is matching the synthetic dataset's performance for the recognition CNN. This is a strong indication that we are exploiting the full potential of the network, despite the lack of a large real dataset.

Pose Estimation As explained in Section 3, in order to evaluate the performance of the CNNs in pose estimation, the inner-distance [25] in cm between predicted and actual grasping points is considered. The results after aggregating over the 180 views for each pose, are presented in CNN_{RD2} column of Table III for the real dataset, whereas CNN_{SD} column contains the results for the synthetic dataset. Single view results are also provided in parenthesis. For comparison, the distances in the related works of [11] and [12] are presented in the second and third columns respectively. These distances as explained in the aforementioned works are also aggregated over 90 and 300 views, respectively, whereas no results on the single view distances were documented. N/A denotes 'not available', since instead of towels in these works shorts have been considered. Even our single view results outperform the aggregated results of the related methods.

However, caution is needed when comparing the results in Table III, since different datasets have been used, whereas training in [11] and [12] is based mostly on simulated 3D models of the garments. Since 3D models are available, in these works distance refers to geodesic distance, approximated in our work by inner-distance on 2D models. Another important parameter is the size of the clothes used for testing. As reported in [11] and [12] the maximum distance between any pair of grasping points (poses), is 75 cm for the shirt, and 65 cm for the pants. In our case, we used articles of clothing presenting larger maximum distances, 108 cm for shirts, 112 cm for pants, and 50 cm for towels.

A comparison of the distribution of the pose Error Distance is plotted in Figure 6 for the different garment types. This is a key evaluation metric in [11] and [12] and is also adopted in our work. As illustrated by the graphs in this figure, our method presents improved results for both shirt and pants, even without normalizing according to the different maximum distances. In case of towels, the correct pose is inferred by the aggregated pose classification results for 84% of the poses. Thus, the corresponding distance error is non-zero only for a small fraction of the test data, rendering the plot of the distance error redundant. As in the recognition case, the above results indicate that the CNNs designed using the large synthetic dataset, perform well even when trained with smaller datasets of real garments. In this case, as implied by the results in Table III, the real datasets outperform SD. A possible explanation is that the synthetic models' configuration when hung is more difficult to discriminate due to the selected material properties and simulation process. However, even in that case, the proposed CNN approach manages to produce acceptable error distances for pose estimation.

Testing the complete pipeline In order to test the complete approach, 6 garments that were not used for training have been employed (2 for each category). Each garment has been autonomously grasped (5 times) by the robotic manipulator and its category and pose have been estimated. In all cases, the category of the garment has been correctly identified, whereas a mean Error Distance of 5.3 cm has been reported for pose estimation. A short video demonstrating the complete recognition pipeline is available online[2].

V. CONCLUSION AND FUTURE WORK

In this work, we propose hierarchical CNNs for recognizing the category and pose of hung garments. Depth images of the garments are used as inputs to the nets and discriminant representations are automatically learned. Both real and synthetic data are employed in the design and evaluation of the networks. Experimental results report above state of the art performance in all datasets, whereas weak cross-domain knowledge transfer has been observed. Our future focus will be on extending the synthetic and real datasets to include additional garment types such as T-shirts and shorts and evaluate the method's performance on the extended datasets. We will also explore multi-view architectures, performing

[2] https://www.youtube.com/watch?v=P25ku9KpMVE

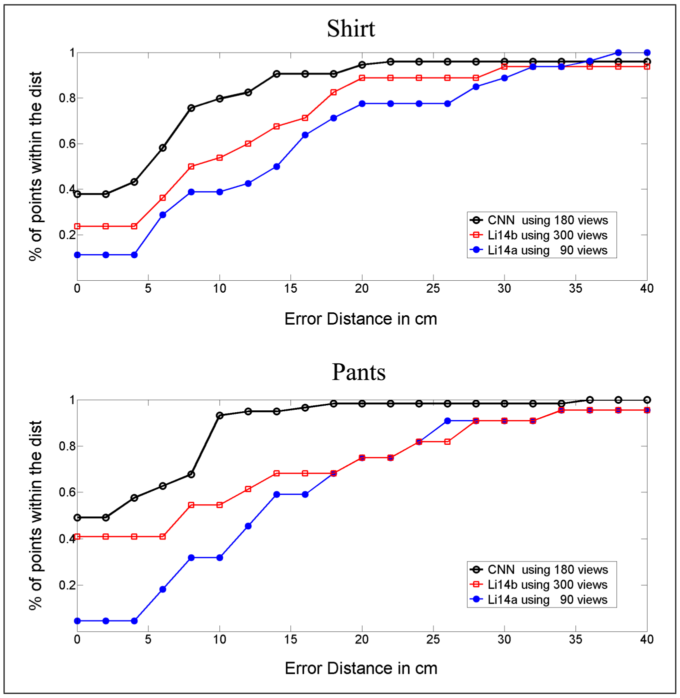

Fig. 6. Error Distance distributions for shirt and pants using RD2 data. Comparison with corresponding distributions in [11] (Li14a) and [12] (Li14b).

fusion of the depth information of different views at the middle layers of the CNNs. Finally, we intend to integrate the method to a broader garment manipulation pipeline, such as picking-up, regrasping, and unfolding.

Acknowledgment

The authors were supported by the EC under project FP7-288553 CloPeMa.

References

[1] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. 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*, ser. CVPR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 943–950.

[2] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3570–3577.

[3] P. Guan, O. Freifeld, and M. J. Black, "A 2d human body model dressed in eigen clothing," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 285–298.

[4] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. International Conference on Computer Vision (ICCV'09)*. IEEE, 2009.

[5] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, "Advances in neural information processing systems 2," D. S. Touretzky, Ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Handwritten Digit Recognition with a Back-propagation Network, pp. 396–404.

[6] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR'04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 97–104.

[7] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 609–616.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013.

[10] Y. Kita, T. Ueshiba, E. S. Neo, and N. Kita, "Clothes state recognition using 3d observed data." in *ICRA*. IEEE, 2009, pp. 1220–1225.

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

[12] Y. Li, Y. Wang, M. Case, S.-F. Chang, and P. K. Allen, "Real-time pose estimation of deformable objects using a volumetric approach," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 1046–1052.

[13] M. Kaneko and M. Kakikura, "Planning Strategy for Putting away Laundry - Isolating and Unfolding Task -," *Symposium on Assembly and Task Planning*, pp. 429–434, 2001.

[14] K. Hamajima and M. Kakikura, "Planning strategy for task of unfolding clothes," *Robotics and Autonomous Systems*, vol. 32, no. August 1999, pp. 145–152, 2000.

[15] F. Osawa, H. Seki, and Y. Kamiya, "Unfolding of Massive Laundry and Classification Types," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, pp. 457–463, 2006.

[16] P. C. Wang, S. Miller, M. Fritz, T. Darrell, and P. Abbeel, "Perception for the manipulation of socks." in *IROS*. IEEE, 2011, pp. 4877–4884.

[17] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA) 2011*, May 2011, pp. 1–8.

[18] S. Miller, J. van den Berg, M. Fritz, T. Darrell, K. Y. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding." *I. J. Robotic Res.*, vol. 31, no. 2, pp. 249–267, 2012.

[19] B. Willimon, S. Birchfield, and I. D. Walker, "Classification of clothing using interactive perception." in *ICRA*. IEEE, 2011, pp. 1862–1868.

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

[21] A. Doumanoglou, A. Kargakos, T.-K. Kim, and S. Malassiotis, "Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning," *ICRA*, 2014.

[22] A. Doumanoglou, T.-K. Kim, X. Zhao, and S. Malassiotis, "Active random forests: An application to autonomous unfolding of clothes," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8693, pp. 644–658.

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

[24] Y. Kita, T. Ueshiba, E. S. Neo, and N. Kita, "A method for handling a specific part of clothing by dual arms," *International Conference on Intelligent Robots and Systems*, vol. 1, pp. 4180–4185, 2009.

[25] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, 2007.

[26] M. J. Thuy-Hong-Loan Le, A. Landini, M. Zoppi, D. Zlatanov, and R. Molfino, "On the development of a specialized flexible gripper for garment handling," *Journal of Automation and Control Engineering Vol*, vol. 1, no. 3, 2013.

[27] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, no. EPFL-CONF-192376, 2011.

[28] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.