Robot-Aided Cloth Classification Using Depth Information and CNNs

Antonio Gabas, Enric Corona, Guillem Alenyà(✉), and Carme Torras

Institut de Robòtica i Informàtica Industrial CSIC-UPC,
C/Llorens i Artigas 4-6, 08028 Barcelona, Spain
galenya@iri.upc.edu

Abstract. We present a system to deal with the problem of classifying garments from a pile of clothes. This system uses a robot arm to extract a garment and show it to a depth camera. Using only depth images of a partial view of the garment as input, a deep convolutional neural network has been trained to classify different types of garments. The robot can rotate the garment along the vertical axis in order to provide different views of the garment to enlarge the prediction confidence and avoid confusions. In addition to obtaining very high classification scores, compared to previous approaches to cloth classification that match the sensed data against a database, our system provides a fast and occlusion-robust solution to the problem.

Keywords: Garment classification · Deep learning · Depth images

1 Introduction

Manipulation of garments, from a highly wrinkled state when they are on a pile to a completely folded state after proper manipulation, is a very challenging task. To enable such manipulation, complex perceptions are required to determine where to grasp, to recognize the type of garment when lying on a table or hanging from the robot hand, to determine the pose or some preferred grasping points, and later on to remove all the wrinkles and perform the necessary folding operations.

The process of isolation of the cloth piece from a pile was identified as one of the first tasks to be solved for laundry manipulation in a pioneer work by Kakikura *et al.* [1,3], where the authors separated and identified three different categories (shirt, pants, and towels) using some ad-hoc rules. The problem of grasping a unique garment from a pile has also been tackled by Monso *et al.* [8]. Where the authors describe the problem of grasping a single garment, as they showed that a naive grasping action executed at the topmost area of a pile can grasp more than one garment at the same time. The authors propose an algorithm where a robot uses a POMDP approach to decide a series of manipulations to increase the probability of grasping only one piece using very simple perceptions.

In this paper, we assume that the robot has already grasped one garment and the objective is to classify it into a category, even if the particular garment has

F.J. Perales and J. Kittler (Eds.): AMDO 2016, LNCS 9756, pp. 16–23, 2016.
DOI: 10.1007/978-3-319-41778-3_2

not been seen before. Similar to our setup, Kita *et al.* have proposed methods to estimate the state of a garment held by a robot. In a seminal work they used monocular views [4], and later they proposed to use depth data obtained using a stereo-rig [5]. Their idea was to pre-compute a database of deformable models and select the best fitting model. Afterwards, they used a second manipulator to change the garment shape and thus increase the recognition accuracy.

More recently, Willimon *et al.* [11] use a stereo rig to classify garments. In their method, the topmost garment of a pile is detected, and grasped using its geometric center. Successive regrasping operations are performed with the end-effector reaching down further until success. Then two images are taken (frontal and side) and the classification process is performed using four basic visual features.

A more elaborated alternative is to use complex feature descriptors. In this direction, Willimon *et al.* [12] explored the use of SIFT and FPFH, and Ramisa *et al.* [10] also compared these descriptors with a new descriptor named FINDDD. This descriptor combines color and depth information to find good candidates for grasping based on a measure of wrinkledness. These works use images of garments lying on a table, and thus are not able to take a second view to disambiguate.

Alternatively, Li *et al.* [6] propose to use a complete 3D model obtained by combining several views using KinectFusion. To obtain this model the robot must grasp and turn completely the garment. Then, 3D features are extracted. Wang *et al.* design specialized features based on coding the distances from the center of the model to its boundaries using cylindrical coordinates. Garments are then identified by matching against a 3-garment database with recognition rates of sweater 85 %, jeans 70 %, and shorts 90 %. Similar to our work, Mariolis *et al.* [7] propose to classify the garments using a Convolutional Neural Network (CNN) as a part of a more general approach. The initial training of the classification CNN is performed in a synthetic dataset and then a real dataset is used for re-training. The authors report an overall success ratio of 89 % over 3 different categories (shirt, pants and towels).

In this paper we propose a method capable of classifying the garment type from the first view, but also able to take advantage of more views in case of confusion by using a robot arm. We propose a Convolutional Neural Network able to perform a fast, robust to occlusions classification and we made the database publicly available. Additionally, we explore the idea of using partial views of the garment. The motivation behind this is that new depth cameras based on ToF sensors are designed to work in closer ranges and some garments, like trousers and unbuttoned shirts, do not fit in the field of view of the camera.

2 Method

The set-up for this system is depicted in Fig. 1, with a pile of clothes placed on a table. We use a Barrett's WAM robot arm for the manipulation of clothing items with a simple gripper, and a Creative Senz3D time of flight (ToF) sensor to capture depth images. ToF cameras (like Kinect ONE, Intel RealSense,

Fig. 1. Setup, including the robot manipulator, the Senz3D ToF camera, and the table where garments are lying.

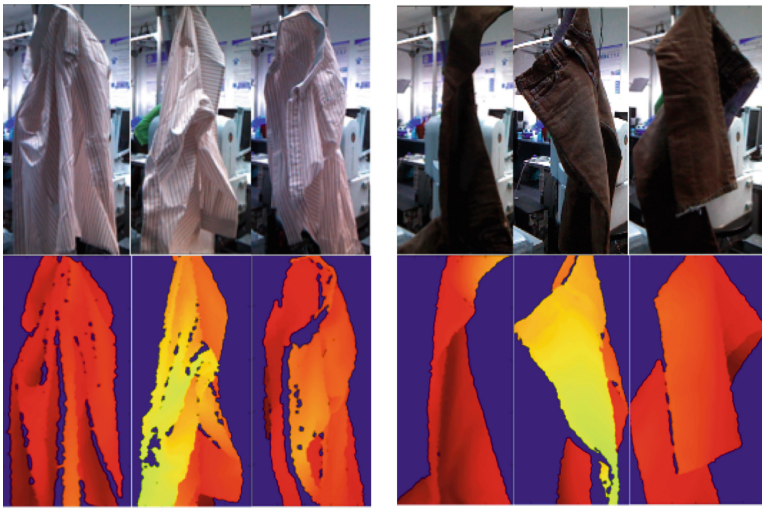
and Softkinetic DS525/Creative Senz3D) are the new depth sensors that are gradually replacing structured light cameras (like Kinect and Xtion) in robotics applications. The main difference is that usually they are designed for closer distances, where the uncertainty of the depth measurements is lower.

We exploit this fact placing the garment close to the camera, at about 70 cm. The images we obtain reproduce with better definition the wrinkles and other details. However, the complete garment does not fit anymore in the image and thus the depth images acquired correspond to parts of the garment. By using only depth data, our system performs the recognition based on the garment's shape and not the texture. This allows the system to generalize between different clothes of the same category.

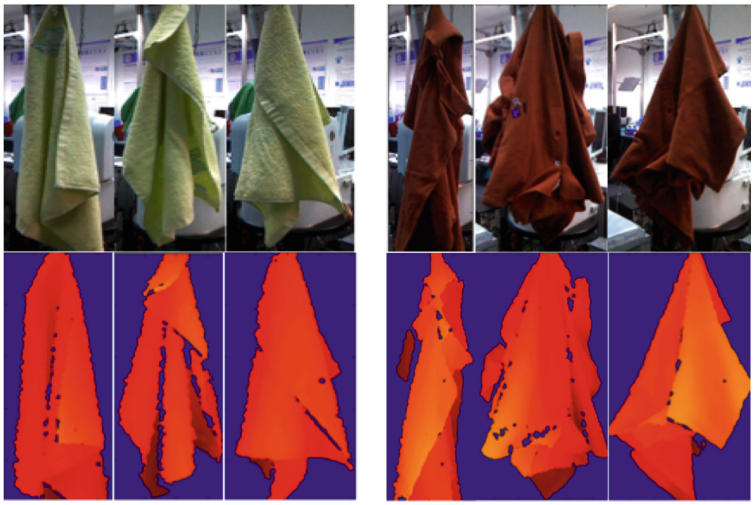
2.1 Dataset Generation

One of the main requirements to train a deep neural network is obtaining a big amount of data. In order to automatize the generation of training data, we programmed the WAM robot arm to repeatedly pick up a garment from a pile by random grasping points and place it in front of the depth camera. Once the robot has centered the garment in the camera image, it starts rotating the cloth item while the sensor captures a total of 12 images per revolution.

By continuous repetition of this operation, we generated a dataset containing 4 types of garments: "Shirt", "Trouser", "Towel" and "Polo" (see Fig. 2). We also detected if the garment felt during the scan. When this occurred, we labelled the empty image as "Nothing". This resulting in the final 5 categories, although the empty cases are not taken into account for the result scoring.



(a) Shirt (b) Trousers



(c) Towel (d) Polo

Fig. 2. Examples of color and depth images from the generated database (observe that the entire garment may not be visible). (Color figure online)

A total of 4272 depth images were obtained from the different categories[1]. 80 % of those images were used to train the network and the remaining 20 % to test it.

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

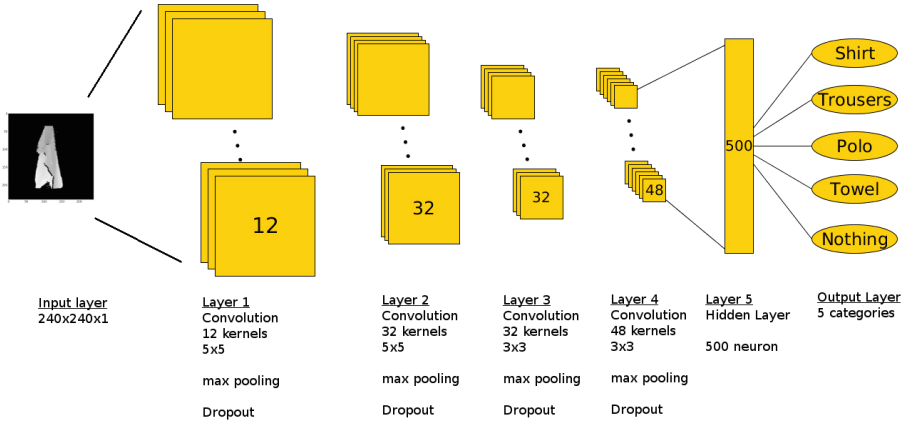


Fig. 3. Architecture of the designed CNN.

2.2 Neural Network Model

We designed a neural network with the architecture depicted in Fig. 3. The model is composed of 5 layers. The first 4 of them are convolutional layers and the last one is fully connected.

The network takes as input a single-channel 240×240 pixel depth image. The first layer consists of a convolution step with 12 kernels of 5×5 pixel followed by a two-fold sub-sampling or max-pooling layer. The second layer consists of 32 kernels of 5×5 pixels plus the max-pooling step. The third layer is equal to the previous one but with kernel size 3×3. The last convolutional layer has 48 3×3 sized kernels and a max-pooling step. Next, the output of the convolutional layer is reshaped as a 1D vector and fed as input to the hidden layer formed by 500 neurons. Finally, the output layer is a softmax layer that outputs the aforementioned 5 classes.

All layers are configured with Rectifier Linear Units (ReLUs), which have proven to be faster than their equivalent tanh units [9]. Also, at the output of each layer, a dropout is performed in order to avoid co-dependences between different nodes [2].

3 Experiments

3.1 Training

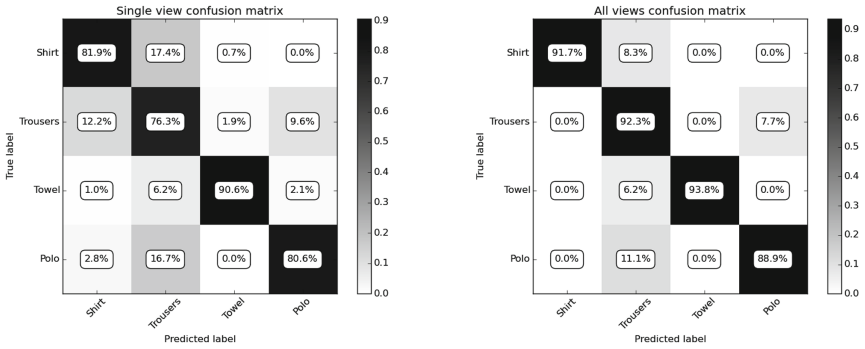
The CNN was trained using stochastic gradient descent with a learning rate of 0.05 and a batch size of 30. The initialization of the network weights is sampled randomly from a uniform distribution in the range [-1/*fan-in*, 1/*fan-in*], where *fan-in* is the number of inputs to a hidden unit (taking into account the number of input feature maps and the size of the receptive fields).

For the prevention of over-fitting we use L2 regularization and dropout: L2 Regularization is a very common form of regularization consisting in penalizing the squared magnitude of all parameters directly in the objective. That is, for every weight ω in the network, we add the term $\frac{1}{2}\lambda w^2$ to the objective, where λ is the regularization strength. We set the value of lambda to 0.0001.

We complement the L2 regularization with the Dropout technique. This method deactivates neurons during training with a probability lower than a given threshold. In this application we obtained good results with a value of 0.3. This prevents co-dependencies between different nodes.

3.2 Results

First, we test the network by feeding to it only one example (i.e. one depth image of a single view of the cloth piece). In this experiment, we obtain a 83 % global recognition rate. In the first column of Table 1 we show the recognition rates for all the cloth types in this experiment. As one would expect, the towel is the most distinctive object and thus is the best classified. The trousers, on the contrary, are the most misclassified ones. We also draw a confusion matrix in Fig. 4a to illustrate what elements are prone to be confused. As mentioned, the trousers are the element that causes more confusion. Besides the case where the trousers are involved, the polo-shirt case is the next that causes more confusion (it gets misclassified 2.8 % of the times). There is almost no confusion between polo and shirt because we used unbuttoned shirts that look quite different.



(a) Single view (b) After 12 views

Fig. 4. Confusion matrices between the different categories.

For the next experiment we take into account all 12 views acquired during the rotation of the garment and select the type with more votes. We assume that similar views will not provide significant new information, so we use the most distant views: for the case of 2 images we use all pairs of opposite views, for the case of 3 images any triplet of equally-rotated and so on. Table 1 shows

Table 1. Recognition rate from 1 to 12 views.

	1	2	3	4	5	6	7	8	9	10	11	12
Shirt	81,9	91	85,4	92,4	87,5	93,8	90,3	95,1	93,8	95,1	91,7	91,7
Trousers	76,3	76,3	77,6	76,9	84,6	82,1	89,1	87,8	90,4	89,1	92,3	92,3
Towel	90,6	91,1	91,1	91,1	91,1	91,7	91,7	93,2	92,7	93,8	93,8	93,8
Polo	80,6	76,9	80,6	78,7	81,5	80,6	86,1	86,1	88,9	88,9	88,9	88,9
All	83	84,7	84,3	85,5	86,8	87,7	89,7	91	91,7	92	92	92

the results of taking into account from 2 to 12 images. As can be seen, the recognition rate is highly improved going from 83 % in the single-view case to a 92 % global score. This big improvement in the recognition is caused by the high recognition increase in the case of the trousers. The trousers are the biggest object in the dataset. Acquiring more views allows the system to disambiguate in complicated cases like partial views of the trousers. The confusion matrix for this scenario is shown in Fig. 4b. As stated, the confusion between items has drastically dropped. The classification in the polo-shirt case has also improved and now we only find misclassifications when the trousers are involved.

The rotation of the garment must be slow in order to avoid oscillation effects. Given that the CNN response is very quick, the cloth movement is the main time-consuming operation. In our tests we obtained up to 12 views of each garment. Other approaches like [7] capture 18 views. We show that after obtaining 9 views results do not improve significantly. The data obtained in these experiments should help in speeding-up the classification process: the robot can stop once enough confidence is reached, or it can decide beforehand the minimum number of views.

4 Conclusions

We have designed a system capable of grabbing a cloth item with a robot arm and use a Deep Convolutional Neural Network to classify its type. Contrary to other approaches, we deal with partial views of garments that cause confusion in the classification. Experimental results show that we can highly improve the classification score by using the robot arm to rotate the garment to obtain more views of it. The obtained results should be later used in a complete system to decide the number of required views and best viewing directions.

As a future work, the designed system can easily be scaled to work with more types of garments with just a few modifications and a bigger dataset. Another interesting topic for future avenues of research is to take the series of captures taken when the robot is rotating the same garment and input them to a Recursive Neural Network. This can help bringing temporal consistency between consecutive individual predictions.

Acknowledgments. This work was partially supported by the EU CHIST-ERA I-DRESS project PCIN-2015-147, by the Spanish Ministry of Economy and Competitiveness under project Robinstruct TIN2014-58178-R, and by the CSIC project TextilRob 201550E028.

References

1. Hamajima, K., Kakikura, M.: Planning strategy for unfolding task of clothes-isolation of clothes from washed mass. In: SICE Annual Conference, pp. 1237–1242 (1996)
2. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv: 1207.0580, pp. 1–18 (2012)
3. Kaneko, M., Kakikura, M.: Planning strategy for putting away laundry-isolating and unfolding task. In: Symposium on Assembly and Task Planning, pp. 429–434 (2001)
4. Kita, Y., Kita, N.: A model-driven method of estimating the state of clothes for manipulating it. In: Workshop on Applications of Computer Vision, pp. 63–69 (2002)
5. Kita, Y., Neo, E.S., Ueshiba, T., Kita, N.: Clothes handling using visual recognition in cooperation with actions. In: International Conferenceon Intelligent Robots and Systems (IROS), pp. 2710–2715 (2010)
6. Li, Y., Wang, Y., Case, M., Chang, S.f., Allen, P.K.: Real-time pose estimation of deformable objects using a volumetric approach. In: International Conference on Intelligent Robots and Systems (IROS), pp. 1046–1052 (2014)
7. Mariolis, I., Peleka, G., Kargakos, A., Malassiotis, S.: Pose and category recognition of highly deformable objects using deep learning. In: International Conference on Advanced Robotics (ICAR), pp. 655–662 (2015)
8. Monsó, P., Alenyà, G., Torras, C.: Pomdp approach to robotized clothes separation. In: International Conference on Intelligent Robots and Systems (IROS), pp. 1324–1329 (2012)
9. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: International Conference on Machine Learning, pp. 807–814. No. 3 (2010)
10. Ramisa, A., Alenyà, G., Moreno-Noguer, F., Torras, C.: FINDDD: A fast 3D descriptor to characterize textiles for robot manipulation. In: International Conference on Intelligent Robots and Systems (IROS), pp. 824–830 (2013)
11. Willimon, B., Birchfield, S., Walker, I.: Classification of clothing using interactive perception. In: International Conference on Robotics and Automation (ICRA), pp. 1862–1868 (2011)
12. Willimon, B., Walker, I., Birchfield, S.: A new approach to clothing classification using mid-level layers. In: International Conference on Robotics and Automation (ICRA), pp. 4271–4278 (2013)

http://www.springer.com/978-3-319-41777-6

Articulated Motion and Deformable Objects
9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings
Perales, F.J.; Kittler, J. (Eds.)
2016, XII, 219 p. 91 illus., Softcover
ISBN: 978-3-319-41777-6