Gripping Positions Selection for Unfolding a Rectangular Cloth Product

Kimitoshi Yamazaki[1]

***Abstract*— This paper describes a method of gripping positions selection from an item of rectangular cloth placed on a table. To select two appropriate gripping positions on the cloth with unarranged shape, we propose a novel method. The proposed method is a development of the previous work of the author and uses a convolutional neural network. One characteristics of the method is to directly compute gripping positions coordinates by matrix calculation using feature vectors extracted from the layer at the final stage of the neural network. We introduce an improved mechanism to estimate the position of clothing hem, and also an improvement on the part to calculate the gripping positions coordinates. The effectiveness of the proposed method was verified using images of actual cloth products.**

I. Introduction

Fabric products are an indispensable part of our daily lives. Therefore, maintenance such as washing and storage is frequently necessary, and we repeat such work every day. Therefore, it is natural that requests for automation of work come out, which is not limited to homes, but also in manufacturing, logistics and many other scenes.

In this paper, we describe a method of gripping positions selection from cloth products which the author considers as one of the important tasks relating to deformable object manipulation. Assume that there is a rectangular cloth with unarranged shape. When picking it up and then aiming to move to a folding action, it is necessary to decide which parts to grasp. What should be noted here is that if it is possible to select two or more positions to be gripped at the same time, it is possible to unfold the cloth quickly, and then it becomes easier to make a prospect for folding afterwards.

Choosing multiple gripping positions at the same time includes a certain difficulty. If the task is to select only one gripping position, there is a high possibility that it will be possible to solve by using a detector for cloth's corner. However, for example, if the task is to unfold a towel by picking up two corners, only applying corner detection is insufficient because there is a possibility to select two corners on the diagonal of the towel. In such a case, we cannot spread it properly when lifting it. That is, it is necessary to select the corner portion in consideration of the connection of the cloth hem. It means that we need to make a comprehensive judgment by combining local information such as a corner portion and global information such as the shape of a cloth product.

*This work was partly supported by NEDO.

[1]Yamazaki is with Faculty of Engineering, Shinshu University, 4-17-1 Wakasato, Nagano, Nagano, Japan. kyamazaki@shinshu-u.ac.jp

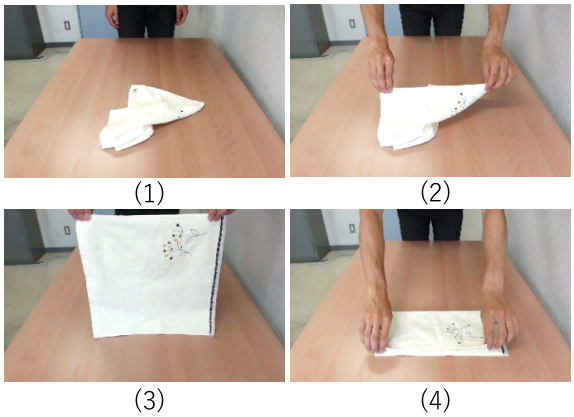

Fig. 1. Unfolding operation using two hands.

In accordance with this idea, the author has proposed several gripping positions selection methods. In [1], we proposed a method using a depth image as input, then estimate gripping positions based on a local representation called hem element. We showed that two appropriate gripping positions can be selected considering both the local shape and the global shape. However, it was necessary that the cloth product was placed with a shape state capable of extracting 3D hem elements. In addition, the accuracy of selecting appropriate holding positions was not yet high. In [2], we proposed a method of selecting gripping positions using a color image. The proposed method obtains four grip position coordinates by matrix calculation from a feature vector obtained from CNN.

In this paper, we extend the method of [2] and propose a novel gripping positions selection method. We improve the accuracy of the selection results which was obtained in the past study. The contributions of this study are as follows.

- We propose a method to properly select multiple gripping positions from a cloth product with an unarranged shape.
- Based on the previous method of inputting one color image, we propose an improvement to enhance the performance by estimating the position of clothing hem.
- We evaluated the proposed method using an actual image set of a rectangular fabric product.

The structure of this paper is as follows. In Section II, we introduce related work and show the difference of our study. In Section III, we discuss the problem setting and our approach. In Section IV, the previous method that is directly related to this study is explained, and improvements and development for the proposed method is explained in Section

V. In Section VI, we describe the experimental results and summarize in Sectin VII.

II. Related work

There have been many cloth manipulation approaches to make a desired shape state. Yuba et al. [3] proposed a method to make corners of a cloth product visible by a predefined action, and then unfolded the cloth by pinch-and-slide motion. In this study, a simple corner detector was used to determine the gripping position. Hamajima et al. [4] achieved an isolation task using color information. The work was taken to pick up the desired cloth product under the premise that each cloth product has a different color, but there was no deep pursuit of grasping position. On the other hand, our aim is to select multiple gripping positions simultaneously from randomly placed cloth products. This facilitates work such as unfolding.

Other studies have also achieved advanced cloth product operation with smarter grasping position selection. Ramisa et al. [5] targeted highly wrinkly clothes and demonstrated an application to detect the collar on crumpled polo shirts as grasp position candidates. Willimon et al. [6] achieved a task of picking a piece of cloth product by one grasping point. Cuén-Rochín et al. [7] proposed an action selection method for handling a planar cloth. In both cases, a cloth product was recognized in the suspended state. Kita et al. [8] proposed a method of matching the model with a 3D point cloud measured using a trinocular stereo camera. The matching process includes a registration to a deformable shape model for the hanging state. Osawa et al. [9] and Maitin-Sphepard et al. [10] succeeded in identifying the type of cloth products by observing the contour and the position of the lower end point while the robot operates cloth products. Doumanoglou et al. [11] performed discrimination and unfolding of cloth products using a three-dimensional range camera. They proposed a method to select the next gripping point from a hanging state for continuous operation. The advantage of these methods is that once the robot picks an cloth item up, it is possible to achieve the manipulation by a relatively high success rate. However, in the above-mentioned studies, detected gripping position is basically a single point. Therefore, it cannot be directly used as a technique for efficiently unfolding cloth placed on a table.

III. Our approach

A rectangular cloth product as the target object is assumed to be placed on a horizontal plane such as a table. The goal of this study is to find appropriate gripping positions from this situation. The gripping positions as mentioned here means several places where cloth products can be well unfolded by pinching and lifting at the points. In the above setting, to detect gripping positions from the cloth product, it is necessary not only to find the position to be grasped locally but also to consider the overall shape state of the cloth product.

Base d on the above, we use a convolutional neural network(CNN) that can output gripping positions as several coordinates. This configuration is based on an assumption that an abstract representation of the overall shape of cloth product includes information on appropriate gripping positions. As explained the above, for unfolding a rectangular cloth, it is necessary to detect two adjacent corners. For this purpose, the overall shape of cloth product should be incorporated as information for making decision.

However, such a CNN is a converter which compresses the overall shape on average, so it is not always possible to extract information suitable for selecting gripping positions. Therefore, we connect an convolutional auto encoder(CAE) for estimating the positions of clothing hem with just before the CNN. It enables to improve the performance of selecting the gripping positions.

There are previous work aimed at calculating gripping positions using CNN. A representative method was proposed by [12]. They succeeded in calculating the position and the angle for a pinching against an object placed on the desk. In that study, the input is a single-color image as the same as ours. Of course, there were studies of the same purpose before then (such as [13]), the work by [12] showed an impressive result as it greatly improved the detection ability by using CNN. However, search and evaluation of a gripping position are performed locally.

From the viewpoint of object detection, recently we can find outstanding methods such as real time conversion [14]. It is possible to detect the position and size of the target object. However, flexible objects such as cloth are not targeted, and it has not been studied to find multiple gripping positions simultaneously from one object.

Another way to use CNN for a gripping position selection is CAE. That is, by inputting a captured image, an image depicted grasping position information is output ([15] and [16]). In such a method, there are cases where a plurality of candidates are obtained. It is also a case that gripping positions are obtained with an ambiguous region. Therefore, it is often necessary to narrow down the candidates and ambiguity. On the other hand, since the approach of this study is a method of directly calculating the coordinates of the gripping positions, no additional processing is required. In addition, above-mentioned conventional studies have basically assumed rigid bodies. It means that these methods implicitly include posture recognition of an object. On the other hand, in this study, CNN is required to be able to recognize the difference in shape condition of deformable objects.

IV. Previous method: Directly calculation approach[2]

A. Method overview

CNN is good at classification problem [17][18]. On the other hand, cloth can take various shape states and it is difficult to make obvious differences on the shape state repertoire. One of the discussion point in this paper is whether the cloth shape state can be classified into several meaningful classes from the viewpoint of selecting the gripping positions.

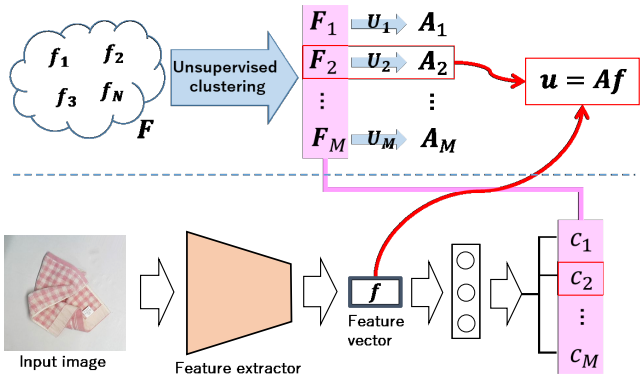

Fig. 2. Structure of the previous method

When a color image is input to the CNN, the position and the shape of cloth are abstracted along deeper shifting the network layer. We hypothesize that this abstraction result contains information suitable for the purpose of estimating the gripping positions appropriately. Therefore, consider improved CNN where the input is a color image and the output is the gripping positions coordinates. In this method, it is necessary to devise such that the range of the output value of CNN does not differ greatly from the scale of image coordinates. Therefore, image coordinates are normalized with the width of the image as a reference. Then, Rectified Linear Unit(ReLU) is introduced in the output layer of the network. As a result, the value of the output becomes within the range of [0.0, 1.0].

Fig. 2 shows the structure of the method. The feature vector $\mathbf{f}$ is obtained from the converter, and a set of gripping position coordinates $\mathbf{u}$ are calculated using it. That is, according to the configuration of ordinary CNN, the convolution layers, the pooling layers and fully-connected layers are utilized to sufficiently abstract image information. After that, the output from that is input to the coordinates calculator, and a set of coordinates are obtained. From the next section, we propose and verify methods based on these items.

B. Selective gripping positions calculation based on clustering result

In previous study [2], we use a structure shown in Fig. 2. As already explained, a color image passes through the feature extractor and the feature vector $\mathbf{f}$ is calculated. Note that another shallow network that recieves the $\mathbf{f}$, and it outputs a classification result. On the other hand, the extracted $\mathbf{f}$ is also used for gripping positions caluclation. By inputting $\mathbf{f}$ to another function obtained beforehand from training data, the vector $\mathbf{u}$ that arranges image coordinates $\mathbf{u}_i = (u_i, v_i)$ with required number is calculated.

For the latter function for gripping positions calculation, the following linear equation is used:

$$\mathbf{A}\mathbf{f} = \mathbf{u}. \tag{1}$$

Hereinafter, the matrix $\mathbf{A}$ is called a transformation matrix. This matrix is obtained beforehand from training data set. The calculation method is as follows. First, N images of cloth products with variously arranged shape are collected as training data. For each image, the appropriate gripping positions $\mathbf{u}_k (k = 1, \cdots, N)$ are selected manually. Then, using CNN, the feature vector $\mathbf{f}_k (k = 1, \cdots, N)$ for each image is calculated. From them, two matrices are generated: A matrix $\mathbf{F}$ composed by vertically arranging N vectors after transposing, a matrix $\mathbf{U}$ composed by vertically arranging vectors $\mathbf{u}_k$ presenting gripping positions coordinates.

$$\mathbf{A}^T = \mathbf{F}^+\mathbf{U}, \tag{2}$$

where $*^T$ indicates transposition. $\mathbf{F}^+$ is the pseudo matrix of $\mathbf{F}$.

C. Improvement of shape classifier

The previous method associates the gripping positions calculation with the classification function. We focus on the effectiveness of considering the overall shape of cloth products, so here we introduce a classifier according to cloth shape.

Various ways are conceivable about the definition of the shape class of the cloth product. As a simple method, unsupervised clustering is applied to training images of cloth products. By assigning the same label to the images in the same cluster, it can be used as a criterion for classification. If the data to be input to this unsupervised clustering is the feature vector $\mathbf{f}$ obtained from the converter, the above procedure can be easily executed. This method is based on the expectation that network weights that are convenient for gripping positions selection are calculated from the class obtained based on the distances in the feature space.

D. Selection of transformation matrix by shape classification result

In IV-B, we explained how to calculate one transformation matrix from all training data. However, since the transformation matrix determined in this way has a large size and contains various shape information, the accuracy of the gripping positions selection cannot be expected in some cases. Therefore, in conjunction with the classification result of the shape class, we will improve the accuracy and reduce the amount of calculation.

Calculation of the gripping positions by Eq. (1) is the same, but processing to select the revised transformation matrix is incorporated in advance. The construction method of this pipeline is as follows. When constructing a shape classifier, we set the number of clusters as M, and divide the training data set by unsupervised clustering. At this time, instead of calculating one transformation matrix from all the training data, we calculate $\mathbf{A}_i$ using only the feature vectors included in each cluster $c_i (i = 1, 2, \cdots, M)$. That is, the same number of transformation matrices as the number of clusters (the number of classes to be classified) is prepared.

When input data is given, the result of shape classification is acquired first. If its output is c_i, gripping positions selection by Eq. (1) is done using $\mathbf{A}_i$. In this method, if the shape classification fails, the estimation accuracy might decrease.

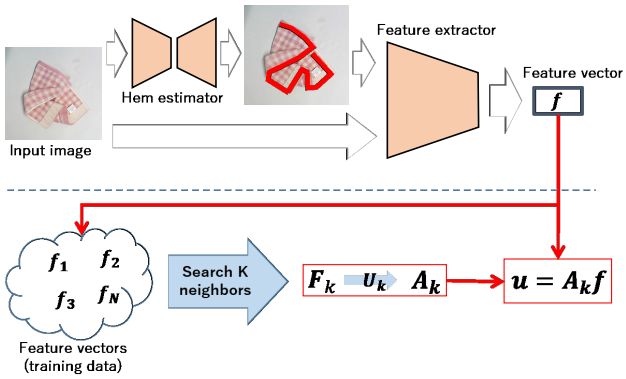

Fig. 3. Structure of the proposed method

However if the classification is successful, the positional accuracy of the gripping positions selection can be greatly increased.

V. The Proposed Method

A. Extension of previous method

The proposed method in this paper is based on the direct method explained in Section IV. Fig. 3 shows the concept of the proposed method. The structure is essentially the same with the above-mentioned directly calculation. However, there are mainly two expansions:

- Using a color image as input data is the same as before. However, the structure of the network immediately after the input is divided into two, and one is made a part to predict the position of fabric hems. After the layer, an intermediate input layer that receives a color image and an estimated hem image is placed. Using this structure, abstract information of the overall cloth shape is obtained.
- Record all feature vectors of the training data. Then, when the output feature vector for the input image is obtained, k vectors close to the input vector are selected on the spot. A transformation matrix is generated by using these vectors, and the gripping positions coordinates are calculated.

The following subsections are explained about these items.

B. The estimation of hem parts

When cloth products are picked up and unfolded, appropriate positions to be gripped are limited. If it is intended to unfold a rectangular cloth, it is desirable to limit the gripping positions to corner parts. However, as described in Section III, it is insufficient to simply find the corners. Therefore, the position of the hem is estimated beforehand as additional information, and it is used for determining the gripping positions.

Estimation of the position of fabric hem is performed using the principle of the de-noising auto encoder. That is, a CNN having the structure of the auto encoder is constituted, learning is performed by inputting a color image and outputting a hem image of the same size as the color image. Hem image here means a mask image created by tracing the hem part of the color image by manual. After learning, when a color image is input to this CAE, an image is obtained. In the image, a high probability density is given to a pixel considered as a part of hem.

When the above was implemented by CAE, it was found that estimation accuracy is higher when inputting after dividing original images into partial images rather than inputting the images as they are. Therefore, we cropoed a color image and a hem image with a size smaller than the size of an original image and learned the CAE. Of course, the input and output of the CAE after learning is a small size image. Therefore, image reconstruction is performed in the following procedure.

First, a partial image with a horizontal width of w and a vertical width of h is cut out from the input image by a stride s. Here, let $\{w, h\} > s$. Each partial image is input to the CAE, and the estimation result is obtained for each pixel in the image. Then the result is mapped it to a resultant image created with the same size as an original image. Each pixel value of the result image is initialized to 0, and the value of the estimation result is added to it. By the above condition $\{w, h\} > s$, the estimated value is added to one pixel of the result image multiple times. This process is performed on all the partial images to obtain the final result image.

C. Calculation of gripping positions

In the proposed method, a number of color images photographing a cloth product is given as training data. As one of the outputs noticing in this paper, one feature vector is obtained from one image. We consider that this feature vector expresses the overall shape of the cloth product. In addition, if some feature vectors are similar, we consider that gripping positions specified there is also similar.

Based on the above assumption, gripping positions coordinates are calculated using $K-$Nearest Neighbor method. That is, K feature vectors closest to the feature vector of the input are extracted from the training data, and a transformation matrix is composed of these vectors. This eliminates the need for prior clustering, which was required for the previous method. Only K is an adjustable parameter.

VI. Experiments

A. Settings and Dataset

A horizontal table was prepared, and a camera was installed at about 700mm just above the table so that the table top could be taken from directly above. The camera module FCB-M made by Sony was used for the camera. A color image with a resolution of 640×480 was captured. In this setting, one pixel on the image corresponds to approximately 1mm.

A cloth product sized 340×340mm was used. 275 images were taken using the cloth. Examples of this dataset is shown in Fig. 4. These were photographed after throwing down cloth products onto the table and the shape conditions were completely random and complicated. Then, two appropriate gripping positions and hem area were manually given to

Fig. 4. Examples of dataset consisting of 275 images

these images, and the image coordinates were recorded in association with the image. In learning process, each image was rotated by 60 degrees in the range of $\pm$ 360 degrees and parallelly shifted by 60 pixels. Then, 15,000 training data were generated. On the other hand, another image different from the training data was prepared and padded by the same method. The number of images of the test data was about 100.

First of all, learning of a CAE for estimating hem parts was performed as advance preparation for gripping positions selection. The size of the input image was 32$\times$32 pixels. About the structure of the network, as a result of trial and error, the number of layers of the encoder section and the decoder section were set to three. The setting of the encoder section in order from the input side is as follows: (number of channels, convolution kernel size, stride) = (8, 9, 2), (16, 7, 2), (16, 5, 2). In each layer, activation was performed by LeRU. On the output side this order was reversed. Adam[19] was used as a learning algorithm, and the coefficient is 1.0×10^{-5}. In the learning process, epoch was 100, batch number was 64. The time required for this learning was about four hours. A computer used in this experiment was Core i7 3.5 GHz CPU, 32 GB memory, and Geforce GTX TITAN 12GB GPU.

Meanwhile, VGG19 proposed by [18] was adopted to extract a feature vector from a color image. A feature vector was extracted from the final hidden layer of the VGG net. That is, the dimension of the vector $\mathbf{f}$ was 4096. As the initial value of the weight, the learning result by the ILSVRC image data set [20] was used. However, as shown in Fig. 3, on the network constructed in this study, its structure is different from the original VGG19. Therefore, for the newly added part, the initial weight was randomly given in the range of 0.0 to 1.0, and appropriate weight was acquired by learning. In this learning, the input was a color images and the output was a pair of gripping positions coordinates. Adam was adopted as optimization method, and the coefficient λ was set to 1.0×10^{-6}. It took two days to learn with epoch 200.

B. Experimental results

Figure 5 is an example of the result of estimating the part of clothing hem using a CAE. First, as shown in the upper left color image and the upper right mask image, a number of color images and hand-written hem images were prepared. Partial images were extracted from them and used as training data for learning of the CAE. Images shown in the lower row are the estimation result of the border obtained by the method described in section V-B. There are two patterns; the

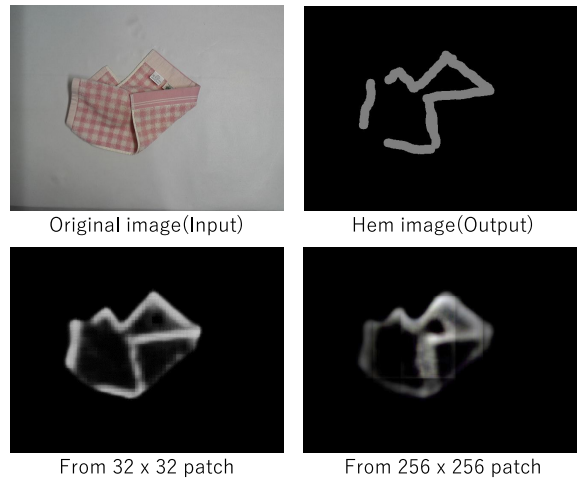

Fig. 5. Examples of hem estimation.

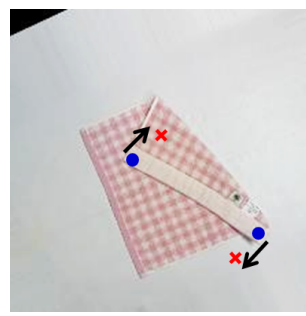

Fig. 6. Definition of error. Difference vectors from the given grasp points are regarded as errors.

size of partial image is 32$\times$32 pixel or 256$\times$256 pixel. It can be seen that learning by the 32$\times$32 size partial image can accurately and clearly estimate the hem part. Note that the folds of the cloth are not recognized as hem part. In the case that 128$\times$128 pixel sized was also tried, and the estimation result was roughly in the middle between the two.

Next, this learning result was incorporated into the structure of Fig. 2, and the selection of the gripping positions was verified. In this proof, two gripping positions were selected from one color image. In evaluating the experimental results, the definition of the error in this experiment was defined as shown in Fig. 6[2]. It is an absolute difference between selected coordinates and manually given coordinates, shown as a black arrow in the figure. If the error value is near to $(0, 0)$, it means that desired gripping positions are obtained. Table I show the error between a selected gripping position and ground truth. Standard deviation of the error at the first one of the two calculated coordinates is shown.

Table 1 summarizes the error of the gripping positions obtained by the proposed method and related method. As explained before, under the conditions of this experiment, 1 pixel corresponds to approximately 1 mm. The top four lines are the result of the proposed method described in this paper, and they were examined by changing the number of k in the

TABLE I

RESULTS OF GRIPPING POSITIONS SELECTION USING ACTUAL IMAGES

	error[pixel]
Ours(k=1)	31.5543
Ours(k=10)	29.1255
Ours(k=30)	31.6707
Ours(k=50)	35.0561
50 clusters	43.3992
100 clusters	39.4641
500 clusters	38.6019
Direct calculation[2]	41.2355
VGG19	42.1070

k-nearest neighbor. The fifth to seventh lines are the results of the proposed method in which only the part to input the estimation of clothing hem is incorporated in the previous method. The numbers 50, 100, and 500 mean number of clusters. The eighth line is the result of the previous method. The ninth line is the result of modified VGG19 network learned by ILSVRC2012 dataset. It was modified so that coordinate values were obtained from the final layer. Transfer learning was performed using the data set shown in this section. For the eighth and ninth lines, padding of data was done more, and the number of data was finally 45000.

From this table, it can be seen that accuracy was slightly improved by incorporating the estimation of clothing hem. Furthermore, by introducing the k nearest neighbor method, accuracy improvement of about 10 mm was seen. That is, the estimation error decreased by 25%. As the best result was obtained when k = 10, it turned out that it is not always true if the k is small. Here, in the case of $k = 1$, the transformation matrix $\mathbf{A}$ was not calculated. The image coordinates corresponding to the closest feature vector were directly used as the estimation result of the gripping positions.

In the table, VGG19 shown in the bottom line means that it constructs a network in which the output from the final layer of original VGG19 becomes gripping point coordinates. Transfer learning was applied to the network, then the performance was verified. The reason for including this evaluation in the experiment is that we can directly know the ability to select the gripping positions of the network. It can be seen that the accuracy was lower than our method.

VII. CONCLUSION

In this paper, we proposed a method of gripping positions selection from an item of rectangular cloth placed on a table. To select two appropriate gripping positions on the cloth with unarranged shape, we introduced a novel CNN structure. One characteristics of the method is to directly compute gripping positions coordinates by matrix calculation using feature vectors extracted from the network. We introduce an improved mechanism to estimate the position of clothing hem, and also an improvement on the part to calculate the gripping positions coordinates. Experimental results showed that the proposed method enables to find gripping points with about 30 mm error in average. This corresponds to an improvement of accuracy of 25%.

Future work, we aim for further improvement in accuracy. Moreover, we apply this method to other types of cloth products, and then implement picking up an item of cloth by actual robots.

REFERENCES

[1] K. Yamazaki: "Grasping Point Selection on an Item of Crumpled Clothing Based on Relational Shape Description," in Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems, ISBN: 978-1-4799-6934-0, 2014.

[2] K. Yamazaki: "Selection of Grasp Points of Cloth Product on a Table Based on Shape Classification Feature," in Proc. of IEEE International Conference on Information and Automation, pp. 136 — 141, 2017.

[3] H. Yuba, S. Arnold and K. Yamazaki: "Unfolding of a Rectangular Cloth Based on Action Selection Depending on Recognition Uncertainty," in Proc. of IEEE/SICE Int'l Symp. on System Integration, pp. 623 - 628, 2015.

[4] K. Hamajima and M. Kakikura: "Planning Strategy for Unfolding Task of Clothes – Isolation of clothes from washed mass –," in Proc. of Int'l. Conf. on Robots and Systems, pp. 1237 – 1242, 2000.

[5] A. Ramisa et al.: "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in Proc. of IEEE Conference on Robotics and Automation, pp. 1703-1708, 2012.

[6] B. Willimon, S. Birchfleld, I. Walker: "Model for Unfolding Laundry using Interactive Perception," in Proc. of IEEE Int'l Conf. on Intelligent Robots and Systems pp. 4871 - 4876, 2011.

[7] S. Cuén-Rochín, J. Andrade-Cetto and C. Torras: "Action Selection for Robotic Manipulation of Deformable Planar Objects," in Proc. of Frontier Science Conference Series for Young Researchers: Experimental Cognitive Robotics, pp. 1 – 6, 2008.

[8] Y. Kita, F. Saito and N. Kita: "A deformable model driven method for handling clothes," in Proc. of Int.l Conf. on Pattern Recognition, Vol.4, pp. 3889 – 3895, 2004.

[9] F. Osawa, H. Seki, and Y. Kamiya: "Unfolding of Massive Laundry and Classification Types by Dual Manipulator," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.11 No.5, pp. 457 – 463 , 2007.

[10] J. Maitin-Sphepard et al.: "Cloth Grasp Point Detection based on Multiple-View Geometric Cues with Application to Robotic Towel Folding," Int'l. Conf. on Robotics and Automation, pp.2308 – 2315, 2010.

[11] A. Doumanoglou, A. Kargakos, T. Kim, S. Malassiotis: "Autonomous Active Recognition and Unfolding of Clothes using Random Decision Forests and Probabilistic Planning," in Proc. of Int'l Conf. on Robotics and Automation, pp. 987 - 993, 2014.

[12] Ian Lenz, Honglak Lee, Ashutosh Saxena: "Deep Learning for Detecting Robotic Grasps," International Journal of Robotics Research (IJRR), vol. 34, no. 4-5, pp. 705-724, 2015.

[13] A. Saxena et al.: "Learning to Grasp Novel Objects Using Vision," Experimental Robotics. Springer Tracts in Advanced Robotics, pp.33 - 42, 2006.

[14] J. Redmon and A. Angelova: "Real-Time Grasp Detection Using Convolutional Neural Networks," in Proc. of IEEE Int'l Conf. on Robotics and Automation, pp. 1316 - 1322, 2015.

[15] A. Nguyen et al.: "Detecting Object Affordances with Convolutional Neural Networks," in Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems, pp. 2765 - 2770, 2016.

[16] S. Levine et al. "End-to-end training of deep visuomotor policies," The Journal of Machine Learning Research Volume 17, Issue 1, pp. 1334-1373, 2016.

[17] Alex Krizhevsky and Sutskever, Ilya and Geoffrey E. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems 25, pp. 1097–1105, 2012.

[18] K. Simonyan and A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Vision and Pattern Recognition, arXiv, 1409.1556, 2014.

[19] D. P. Kingma, J. Ba: "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.

[20] ILSVRC2012 image dataset: http://www.image-net.org/challenges/LSVRC/2012/(viewed 2018/1/10)