Recognizing clothing states using 3D data observed from multiple directions*

Yasuyo Kita, Toshio Ueshiba, Fumio Kanehiro and Nobuyuki Kita

Abstract— In this paper, we propose a method of recognizing the state of a clothing item by using three-dimensional(3D) data observed from multiple directions in an integrated manner. The situation dealt with in this paper is that a clothing item is observed from different directions by rotating it along a vertical axis. First, sets of 3D points obtained from each observation are integrated on a depth buffer image which lies on the side of a cylinder containing the item (CZ buffer image). Then, CZ buffer is expanded into a new depth buffer image, whose horizontal axis is akin to geodesic distance on the clothing surface (EZ buffer image). As a result, the region where 3D points are stored in EZ buffer images approximates " a view of flattened surface" of the item, which is stable regardless of the variation in 3D shape of the item as far as the item is held at the same position. Experimental results using both synthetic images and actually observed images demonstrated that the similarity of regions in EZ buffer images is an effective measure for recognizing clothing states.

I. INTRODUCTION

As importance of home and rehabilitation robots is increasing in an aging society, robots are required to handle a wide variety of common objects including clothing. For handling a clothing item naturally, transferring the item from one hand to the other in air is one of essential function. For executing this automatically, it is important to visually find a part to grasp (target part) from the item that is held by one hand. When the target part is very characteristic, like the corner of towels, the parts can be searched using relatively simple feature detection [1]. Even in the case, however, just grasping the target part without considering total configuration of the item may produce undesirable twisting of the item. Therefore, "recognition of total clothing state", that is the recognition of not only its geometrical shape, but also where each part of the item is in the shape is indispensable.

Some studies take strategy of first unfolding clothing item on a desk to recognize the item without self occlusion [2][3]. This strategy requires a big-sized table, that can be accepted for factory environments but is not suitable in an ordinary daily life. Therefore, we take the strategy to recognize the item hanging in air in a similar way to that we human do. We start from the situation where the item is placed in an arbitrary shape on a desk. Although a method for picking up a characteristic part like collar of shirts first using features

common to various shirts[4], we assume that an arbitrary part is picked up first since clothing items do not always have characteristic parts, and, even in the case it has, the part may be unobserved due to (self-) occlusion.

Kaneko et al. [5] proposed a method that recognizes the clothing state by comparing the contour features(e.g., curvature, length-ratio) of observed data with models under the condition that clothing items are held at two points in air. The state of clothes held at two points, however, has tremendous variations, because of the high number of two-point combinations. Osawa et al. [6] and Cusumano-Towner et al. [7] followed this approach while decreasing the variations by repeating re-grasping of the lowest point of clothing items. It, however, requires unnecessary actions which causes delays in achieving the goal state.

Kita et al.[8][9] proposed a method to recognize the state of a clothing item which is held by one hand under the assumption that the clothing item is not very soft. The method recognizes the clothing state by comparing the observed 3D data with possible shapes which are obtained in advance using simulation of physical deformation of the item. By strategically restricting the variation in its 3D shape using precedent actions, this process can be done with a practical number of representative shapes[10].

In these studies, one observation is used for recognizing the clothing state. Sometimes, however, 3D data obtained from just one view is not enough to determine the state, especially when only a small part of the clothing item is observed in the view. To obtain more 3D data of the item for decreasing ambiguity, it is effective to observe the item from different directions by rotating the item around the vertical axis as shown in Fig. 1(a),(b) and (c). Although Maitin-Shepard et al. [1] also used multiple views obtained while rotating a towel, they did so to search each observation for a characteristic point (concretely, the corner of the towel). On the other hand, for the purpose of recognizing total configuration of the item, we propose to use all 3D data obtained from different directions in an integrated manner.

There are many works on the integration of multiple views of rigid objects[11]. Cylindrical mapping that projects 3D points on one cylindrical image plane (ex. [12]) and records the distance of the points from the axis of the cylinder at the projected image coordinates is one of proper ways to integrate 3D points from different observations. However, the representation of the 3D data using such cylindrical depth image does not directly offer us a good way to understand the clothing state of the data.

In this paper, we propose a new description for 3D points
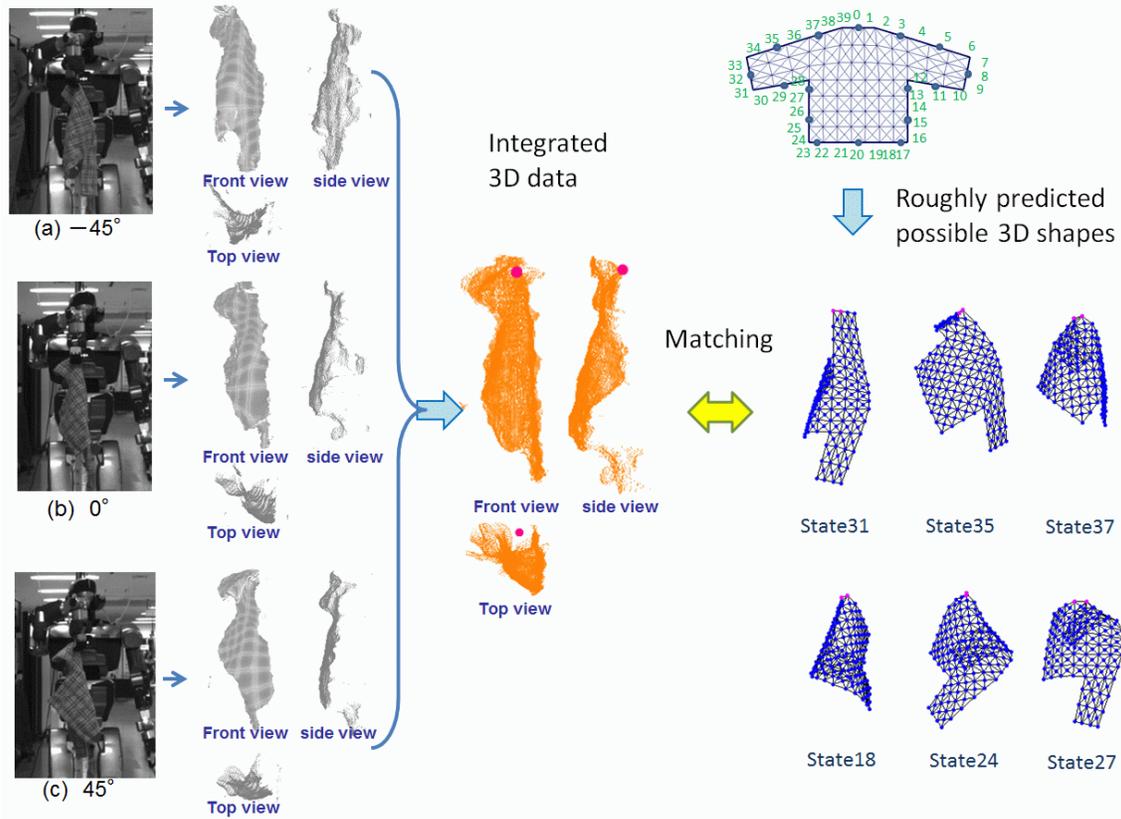
Fig. 1. Purpose of this study: matching of integrated observed 3D data and model shapes

on the clothing surface which can give a kind of "view of the flattened surface" by extending cylindrical depth buffer image. From now, In Section II, background of this research and assumptions used in this paper are briefly explained. In Section III, new description developed in this study and how to use it for recognition of clothing states are explained. In Section IV, Experimental results are shown and discussed.

II. BACKGROUND OF THIS RESEARCH

Suppose that, at first, a clothing item is placed in an arbitrary shape. A humanoid robot, HRP-2 [13], picks it up by one hand and grasp any point along its rim by the other hand [10]. Then, the item is set in a shape convex toward the stereo camera system [14] which is located in front of the humanoid. The state of the item after these precedent actions can be restricted into ones held at its rim. We classified them into "State 1," "State 2," ...,"State n," according to the position at which the item is held, by dividing the rim into n segments. In Fig. 1, we can see examples of some states in the case of a pullover (ex. State 31 and so on), where the rim is divided into $n = 40$ segments. A representative 3D shape at each state, that we call as "model shape" from now, is calculated by simulating the physical deformation of the item under the following assumptions:

1. The clothing items are not very soft and no drapes appear when it is held at a point in air.

2. Clothing type(e.g. pullover, trousers, and so on) and approximate size are given so that its deformable model can be built in advance from a canonical model common to each clothing type.

In addition, in many cases, we can use a simplified model by assuming the front and back sides of the clothing are not separated like a pullover model as shown in Fig. 1.

Concerning the number of observations for recognizing clothing items, smaller is better to shorten execution time. Our stereo camera system can calculate 3D data of the surface the normal of which is within $45 \sim 60$ degrees around the view direction. Considering that the normal directions of most of surface of the item are within 100 degrees from the view direction under the condition that the item is set in a shape convex toward the camera, we observed the item three times while rotating it by -45, 0, 45 degrees in the experiments of this paper.

The recognition of the state of the observed item is performed in two-stages:

I. A state whose representative shape is most consistent with the observed shape is selected for estimating at which point the item is held.

II. The clothing model in the representative shape of the state is deformed so as to fit the observed 3D data better by using virtual attractive force from the observed points to the model vertices.
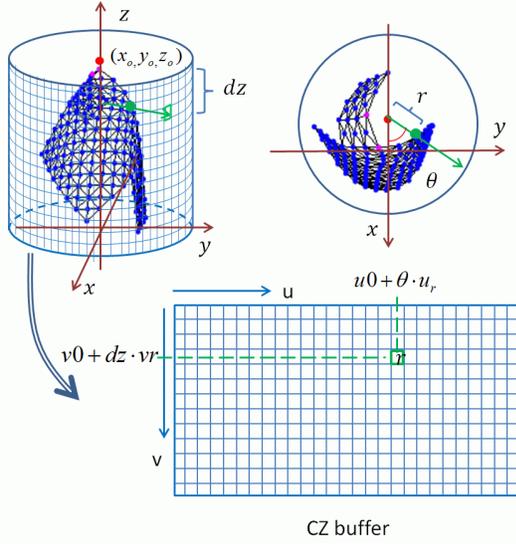
Fig. 2. Transformation of 3D points into a Cylindrical depth buffer (CZ buffer)

Fig. 3. Integration of CZ buffers which are taken while rotating the item along a vertical axis

III. DESCRIPTION OF OBSERVED 3D DATA FOR MATCHING WITH MODEL SHAPE

A. Integration of 3D data using Cylindrical depth buffer image

To integrate 3D points from different observations, we adopt a strategy to project them on one depth buffer image which is set at the side of a cylinder containing the clothing item as shown in Fig. 2. By recording the Euclidean distance of each 3D point from the cylinder axis, r, in the pixel at its projected position (u, v), 3D point cloud is mapped onto this depth image. From now, we call this cylindrical depth buffer image as a cylindrical Z buffer and abbreviate it to CZ buffer. Concretely, a 3D point $P_i(x_i, y_i, z_i)$ is recorded at the coordinates of CZ buffer, CZ[v][u], whose cylindrical axis is the vertical line through (x_o, y_o, z_o) as follows:

$$CZ[v][u] = \sqrt{dx^2 + dy^2} \qquad (1)$$

$$u = round[\theta * u_r + u0]$$
$$v = round[dz * v_r + v0]$$
$$\theta = \arctan{(dy/dx)}$$

$$dx = x_i - x_o, dy = y_i - y_o, dz = z_i - z_o,$$

where, $u0(v0)$ and $u_r(v_r)$ are parameters to determine the center and the resolution of the $u(v)$ axis of CZ buffer respectively; $round[x]$ is the function that rounds x to the nearest integer. When there are more than one point which are projected onto the same coordinates of CZ buffer, the farthest point from the axis of the cylinder is recorded. The purpose of this selection is to record the closest surface to the camera while ignoring the 3D points observed from the backside of the surface.

Concerning the vertical center axis of the cylinder, (x_o, y_o, z_o), we set it at the position away in the minus x direction by l_x from the position where the item is held.

The reason for setting this deviation is to avoid the situation that many observed points come just on or fairly close to the center line, that makes their θ values unstable. In all experiments in this paper, l_x is fixed and set $50mm$.

To integrate the sequence of m observations which are taken while rotating the item by angle θ_k (k=0,1,...,m-1), each CZ buffers, CZ_k can be combined into one CZ buffer by translating them with the variable transformation of $u' = u_k + \theta_k * ur$. Fig. 3 shows an example when the three observations in Fig. 1(a), (b) and (c) are integrated.

B. Expanding of Cylindrical representation for matching with models

Fig. 4(a) shows the representative shape of State 8. On the other hand, Fig. 4(b) shows another possible shape of State 8 which is obtained by simulating the item deformation using a smaller bending resistance parameter of the model. Since such difference in 3D shape between them generally occurs even for the same item, similarity measure which insensitive to these type difference is desirable for the purpose of classifying the observation into one of States.

Fig. 4(d) and (e) are CZ buffer images of the 3D shapes of Fig. 4(a) and (b) respectively. Unfortunately, the coordinate u in CZ buffers is sensitive to how strongly the clothing item is curved. As illustrated in Fig. 4(c), the same vertex of the model has quite different (r, θ) depending on how strongly the item bends.

To obtain representation which is insensitive to, or ideally invariant to, the shape variation of the item held at the same position, we imitate to extend the clothing surface on a flat plane. This is implemented by obtaining a new image, (U,V), whose U coordinate approximates a horizontal geodesic distance on the surface. Concretely, the new Expansion Z buffer image, EZ buffer, EZ[U][V], is calculated from CZ
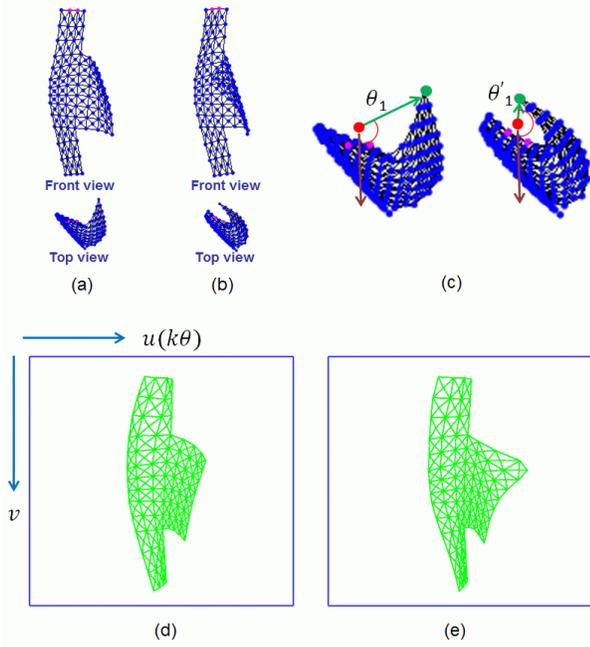
Fig. 4. Difference in Cylindrical Z buffer (CZ buffer) caused by the difference in the amount of curving)

Fig. 5. Transformation of CZ buffer to Expansion Z buffer (EZ buffer)

buffer, CZ[v][u], as follows:

$$EZ[V][U_{i,v}] = CZ[v][u_{i,v}] \qquad (2)$$

$$if \ \ i > 0$$
$$U_{i},v = U_{i-1,v} + round[(r*(u_{i,v} - u_{i-1,v})/u_r) * U_r]$$
$$if \ \ i < 0$$
$$U_{i},v = U_{i+1,v} - round[(r*(u_{i,v} - u_{i+1,v})/u_r) * U_r]$$
$$V = v$$

$$r = (CZ[u_{i-1,v}][v] + CZ[u_{i,v}][v])/2.0$$
$$U_{0,v} = U0 + round[(CZ[u_{0,v}][v] * u_{0,v}/u_r) * U_r]$$

where, $(..., u_{-2,v}, u_{-1,v}, u_{0,v}, u_{1,v}, u_{2,v}, ...)$ is a sequence of the u coordinates of the existing points in the v line of CZ buffer, $u_{0,v}$ is the u coordinate closest to $u0$ in the sequence, $U0$ and U_r are parameters to determine the center and the resolution of the U coordinate of EZ buffer respectively.

Here, horizontal geodesic distance between P_i and P_{i+1} are approximated by $rd\theta$ as illustrated in Fig. 5(a): The reason why this approximation is used instead of ordinary Euclidean distance between the points is to avoid bad influence caused by jaggies of observed surface due to noises and/or quantization.

Fig. 5(b) and (c) are EZ buffer images calculated from Fig. 4(d) and (e) respectively. As we see, region of the 3D data in EZ buffer images are almost coincident with each other irrespective of the difference in the actual 3D shape. This indicates that the similarity of area in EZ buffer image can be used for measuring the closeness to each state.

C. Model fitting using correspondence in EZ buffer image

Once the clothing state (that is, at which part the clothing item is held) is estimated, accurate shape is obtained by deforming the model shape of the state to the shape that fits the observed data better [8]. This can be executed by analogically simulate that the model surface is attracted to corresponding 3D observed points. In the work of [8], we deform the model gradually from the vicinity of the holding position towards the further parts, based on the idea that, at least in the vicinity of the holding position, we can assume that the model is already close to the observed data, so that we can use Euclidean distance as a key for finding correct correspondences between model vertices and the data point. This time, the correspondence can be estimated more directly using EZ buffer, since the observed point corresponding to a model vertex should lie in the vicinity in EZ buffer irrespective of the difference between their 3D positions. Concretely, the observed point which is the closest to each model vertex in EZ buffer image is chose as the corresponding point, while the model is iteratively deformed using similar implementation as described in Section III.B. of the work[8].

IV. EXPERIMENTS

A. Recognition experiments using Synthetic data

To check the stability of regions in EZ buffers, we conducted experiments using synthetic observations which are obtained from physical simulation done while changing the parameters of bending resistance of clothing models. Fig. 6(a) shows an example of three shapes obtained with three different bending resistance, $0.5B_0$, B_0 and $2.0B_0$, where B_0 is the parameter used for calculating representative model shapes. In the experiments, the 3D shapes obtained with the parameters $0.5B_0$ and $2.0B_0$ are used as observed 3D data, while the 3D shapes obtained with the parameters B_0 are used as model shapes. These synthetic data were calculated for seven states, State 0, 3, 5, 8, 13, 17 and 20 ($Sk, k = 1 \sim 7$

Table 1   Area overlap ratio (AR) between synthetic observed data and model 3D shape

	S1-1.0	S2-1.0	S3-1.0	S4-1.0	S5-1.0	S6-1.0	S7-1.0	S8-1.0	S9-1.0	S10-1.0	S11-1.0	S12-1.0
S1-0.5	0.771	0.733	0.765	0.612	0.696	0.741	0.793	0.777	0.691	0.568	0.772	0.78
S2-0.5	0.718	0.921	0.745	0.502	0.736	0.715	0.767	0.76	0.701	0.57	0.685	0.74
S3-0.5	0.62	0.722	0.937	0.647	0.701	0.717	0.7	0.809	0.722	0.727	0.72	0.66
S4-0.5	0.476	0.482	0.609	0.919	0.565	0.677	0.541	0.628	0.558	0.719	0.736	0.543
S5-0.5	0.703	0.751	0.716	0.569	0.939	0.674	0.725	0.741	0.789	0.559	0.745	0.743
S6-0.5	0.655	0.662	0.735	0.667	0.632	0.914	0.66	0.653	0.702	0.67	0.798	0.681
S7-0.5	0.701	0.76	0.696	0.537	0.685	0.713	0.93	0.69	0.686	0.55	0.697	0.743

	S1-1.0	S2-1.0	S3-1.0	S4-1.0	S5-1.0	S6-1.0	S7-1.0	S8-1.0	S9-1.0	S10-1.0	S11-1.0	S12-1.0
S1-2.0	0.892	0.741	0.622	0.44	0.708	0.668	0.712	0.649	0.705	0.448	0.617	0.706
S2-2.0	0.702	0.933	0.696	0.468	0.745	0.689	0.748	0.68	0.731	0.509	0.654	0.721
S3-2.0	0.619	0.66	0.879	0.612	0.695	0.704	0.647	0.811	0.693	0.728	0.695	0.639
S4-2.0	0.491	0.48	0.643	0.907	0.582	0.652	0.489	0.64	0.578	0.73	0.671	0.481
S5-2.0	0.647	0.668	0.648	0.552	0.869	0.62	0.681	0.705	0.737	0.531	0.684	0.718
S6-2.0	0.655	0.662	0.735	0.667	0.632	0.914	0.66	0.653	0.702	0.67	0.798	0.681
S7-2.0	0.756	0.78	0.676	0.495	0.713	0.672	0.928	0.673	0.712	0.498	0.673	0.78

Fig. 6.   Examples of similarity of EZ buffer (with synthetic data): (a) comparison between largely different 3D shapes; (b) case of low AR between corresponding 3D shapes.

in Table 1), and compared with model shapes of the same seven states and its symmetrical states, State 23, 27, 32 and 35 ($Sk, k = 8 \sim 12$).

In all experiments in this paper, area overlap ratio, AR, which is the average of the ratio of overlapped area to model area and the ratio of overlapped area to observed area, is used as the measure of the similarity of regions in EZ buffer images.

In Table 1, AR values between each synthetic data (Sx-0.5 and Sx-2.0) and 12 model shapes are summed up. The highest values of AR for each synthetic data is marked with orange color, while the values larger than 0.95 times of the max value is marked with yellow color. In all experiments except the case of S1-0.5, correct states showed the highest values with a relatively big difference from the second highest. These results indicate the stability of this measure.

To show the influence of large shape difference to this similariy, comparison of S3-0.5 and S3-2.0 is shown in Fig. 6(a), where S3-0.5 is used as model shape, while S3-2.0 as observed one. Observed region, model region and overlapped regions in EZ buffer images are shown with green, red and yellow colors respectively. Despite the large difference between their 3D shapes, the value of AR is 0.872, that is fairly high as compared with the values between mismatched states shown in Table 1.

In the case of s1-0.5, its correct state, s1-1.0 did not show the highest value. The reason of this low AR is that most part of the sleeves do not appear in CZ/EZ buffer image since the axis of cylinder for CZ buffer lies outside of the area surrounded by the clothing surface. However, in this case, there is no sharp peak with many similar values (yellow columns). Therefore, it may be possible that the system automatically excludes this as an ambiguous case.

B. Recognition experiments using 3D data observed by stereo camera system

We conducted the experiments using images of a pullover observed by our trinocular stereo system while changing its holding position along the rim. Pink circles in Fig. 7(a) shows 12 positions tried in the experiments. At each position, three observations are taken from -45, 0, 45 degrees from the view direction of the camera system. In Fig. 7(b), green and red points show the position of the pixels where the observed
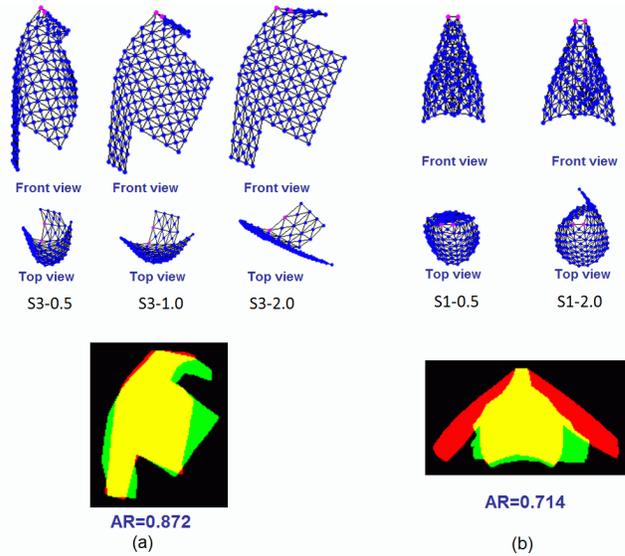
3D points are stored in CZ buffer and EZ buffer respectively. As you see, in most of EZ buffers, distribution of red points look close to "view of flattened surface" of each state. Note that "view of flattened surface" can be lack of sleeve parts depending on the state; as an example, one sleeve is lacked because it is sharply folded on the body part in Data 4 of Fig. 7(b).

By calculating AR between each EZ image with model shapes for 40 states and selecting the state which gives the highest value, the state of each observation, is determined. For eight among 12 data, states were correctly detected. Fig. 8 shows a successful example, the result of Data 9, whose observations from three different directions are shown in Fig. 1. Here, AR of State 34, 0.806, was the maximum. Details of the region similarity are shown in Fig. 8(a). Then, model shapes at states in the vicinity of State 34 are deformed so as to fit the observed 3D data in the way explained in Subsection III-C. Fig. 8(b) shows an example of point to point correspondences between model vertices and observed 3D points, which were found using the EZ buffer. The final decision is made by the best fitted shape. In this case, State 35 shows the best fit as shown by the model superposed on the observed data in Fig. 8(b) and (c).

In two cases, Data 5 and 12, states adjacent to each correct state were selected, like State 37 for correct State 39 of Data 12 as shown in Fig. 9(a). In the two remaining cases, Data 3 and 11, wrong states were selected. Fig. 9(b) shows the result of Data 11. Most of these failures comes from the lack of sleeve regions in observed data caused by sharp folding of sleeves at the shoulder contrary to the corresponding model shapes. As a result, AR of the correct states became low. Since such folding is predictable, one solution to avoid this is to add model shapes with sleeve folded sharply only for the states where such folding can happen.
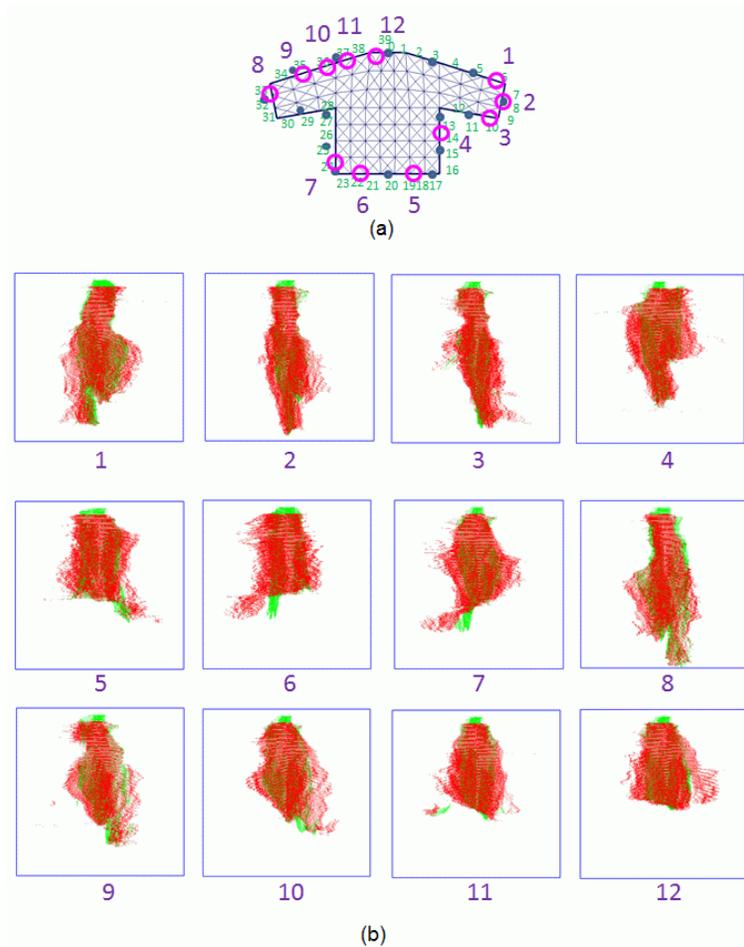
Fig. 7. EZ image calculated from 3D data obtained from stereo cameras: (a) 12 holding positions for experiments (pink circles); (b) Results of CZ (green) and EZ (red) buffer images.
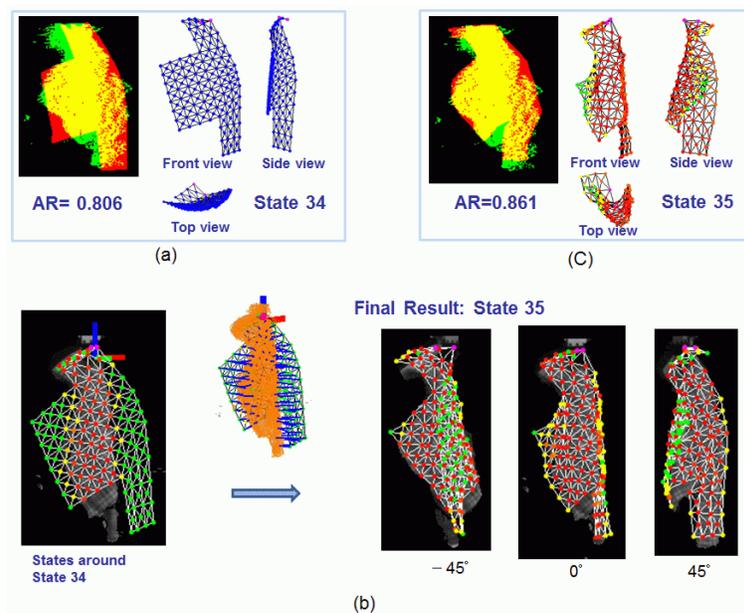
Fig. 8. Experiment results (success case): (a) region similarity with model shape of the corresponding state; (b) process of fitting of the model to the observed data; (c) region similarity after fitting.
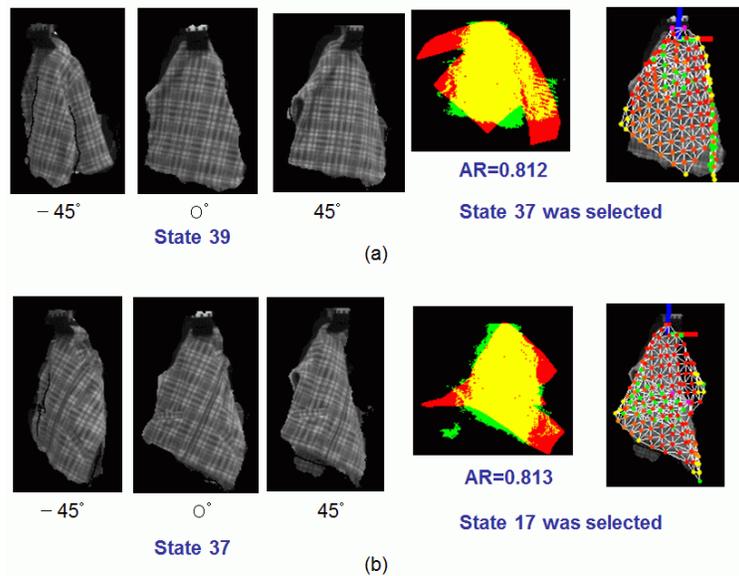
Fig. 9. Experiment results (failure cases): (a) state adjacent to the correct state was selected; (b) wrong state was selected.

V. CONCLUSION

In this paper, we proposed a new Z buffer image for recognizing the state of clothing items using 3D data observed from multiple directions. To understand the state of a clothing item held in air, it is important to know at which position the item is held. Considering that the shapes of the item held at the same position show similar views when it is flattened on a plane, we developed a new depth image buffer, in which the region where 3D points are stored becomes akin to " a view of flattened surface".

For the calculation of the new description, horizontal geodesic distance on the clothing surface is approximated by $rd\theta$ in the cylindrical representation of the surface. This approximation works well to flatten the surface which is relatively smooth. Unfortunately, this can not be used to flatten big drapes. In the first place, however, it is difficult to obtain accurate 3D data of the surface with such big drapes. Therefore, the approach using supportive action that eliminates drapes before observing it should be a proper direction to deal with very soft clothing. This is one of our future subjects.

ACKNOWLEDGMENT

REFERENCES

[1] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei and P. Abbeel:"Cloth Grasp Point Detection based on Multiple-View Geometric Cues with Application to Robotic Towel Folding", In *Proc. of IEEE Int'l Conf. on Robotics and Automation (ICRA '10)*, 2010.

[2] S. Miller, M. Fritz, T. Darrell and P. Abbeel: "Parameterized Shape Models for Clothing", In *Proc. of IEEE Int'l Conf. on Robotics and Automation (ICRA '11)*, pp. 4861–4868, 2011.

[3] B. Willimon, S. Birchfield, and I. Walker: "Model for unfolding laundry using interactive perception", In *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS11)*, pp. 4871–4876, 2011.

[4] A. Ramisa, G. Alenya, F. Moreno-Noguer and C. Torras: "Using depth and appearance features for informed robot grasping of highly wrinkled clothes", In *Proc. of International Conference on Robots and Systems*, pp. 1703–1708, 2012.

[5] M. Kaneko and M. Kakikura: "Planning strategy for putting away laundry –Isolating and unfolding task –", In *Proc. of the 4th IEEE International Symposium on Assembly and Task Planning*, pp. 429–434, 2001.

[6] F. Osawa, H. Seki, and Y. Kamiya: "Unfolding of massive laundry and classification types by dual manipulator", *JACIII*, 11,5, 457–463, 2007.

[7] Marco Cusumano-Towner, Arjun Singh, Stephen Miller, James O'Brien, Pieter Abbeel: "Bringing Clothing into Desired Configurations with Limited Perception", In *Proc. of IEEE Int'l Conf. on Robotics and Automation (ICRA '11)*, 2011.

[8] Y. Kita, T. Ueshiba, E. S. Neo and N. Kita:"Clothes state recognition using 3D observed data", In *Proc. of International Conference on Robotics and Automation*, pp. 1220–1225, 2009.

[9] Y. Kita, T. Ueshiba, E. S. Neo and N. Kita: "A method for handling a specific part of clothes by dual arms", In *Proc. of International Conference on Robots and Systems*, pp. 480–485, 2009.

[10] Y. Kita, F. Kanehiro, T. Ueshiba and N. Kita:"Clothes handling based on recognition by strategic observation", In *Proc. of International Conference on Humanoid*, pp.53–58, 2011.

[11] J. Salvi, C. Matabosch, D. Fofi and J. Forest: "A review of recent range image registration methods with accuracy evaluation", *Image and Vision Computing*, 25, 5, pp. 578–596, 2007.

[12] I. R. Khan, M. Okuda, and S. Takahashi:"Regular 3D mesh reconstruction based on cylindrical mapping", In *Proc. International Conference on Multimedia and Expo*, pp.133–136, 2004.

[13] Kenji Kaneko and Fumio Kanehiro and Shuuji Kajita and Masaru Hirata and Kazuhiko Akachi and Takakatsu Isozumi: "Humanoid Robot HRP-2", In *Proc. of IEEE Int'l Conf. on Robotics and Automation (ICRA '04)*, pp.1083–1090, 2004.

[14] T. Ueshiba: "An Efficient Implementation Technique of Bidirectional Matching for Real-time Trinocular Stereo Vision", In *Proc. of 18th Int. Conf. on Pattern Recognition*, pp.1076–1079, 2006.