Proceedings of the 2004 IEEE
International Conference on Robotics & Automation
New Orleans, LA • April 2004

A deformable model driven visual method for handling clothes

Yasuyo Kita Fuminori Saito Nobuyuki Kita
Intelligent Systems Institute,
National Institute of Advanced Industrial Science and Technology (AIST)
AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki, Japan 305-8568

Abstract- **In this paper, we proposed a deformable model-driven method to obtain the 3D information necessary for handling a clothes by manipulators from observation with stereo cameras. The task considered in this paper is to hold up a target part of clothes (e.g. one shoulder of a pullover) by the second manipulator, when the clothes is held in the air at any point by the first manipulator. First, the method calculates possible 3D shapes of the hanging clothes by simulating the clothes deformation. The 3D shape whose appearance gives the best fit with the observed images is selected as estimation of the current state. Then, based on the estimated shape, the 3D position and normal direction of the part where the second manipulator should hold are calculated. The results of preliminary experiments using actual two manipulators have shown the good potential of the proposed method.**

Index Terms- deformable model, robot vision, clothes handling, manipulation

I. Introduction

The handling of soft objects is attracting increasing attention in the robotics field. However, it is still challenging, though essential, to visually understand the state of largely deformed objects. Although rope handling has been studied [1], in case of dealing with clothes, complex self-occlusion makes it very difficult to understand the state. In the field of computer graphics, sophisticated models for animating cloth deformation have been developed [2]. However, research on automatic recognition of deformed clothes has just started [3][4]. Although it is possible to obtain the detailed 3D information of clothes using elaborate range sensors [5], we aim to extract only the information necessary for a given task from images taken by a simple camera system, as human beings do daily.

Kaneko et. al[3] proposed a method which recognizes the clothes state by comparing the contour features (e. g. curvature, length-ratio) of an observed appearance with ones of model appearances under the situation that the clothes is hanging at the two points. However, the contour features are difficult to robustly extract from real observations and are very sensitive to a slight deformation of clothes. Additionally, its learning processes to obtain the model appearances from actual observations are troublesome. In [4], we proposed a method which recognizes the state of clothes hanging at a point in a model driven way. The method predicts the possible appearances using a deformable model of the clothes and selects one which fits the observed appearance the best. The results of the experiments using a pullover held by a human hand were encouraging. Although the method succeeded in indicating the position of the part to be held next on observed images, the information is two-dimensional image coordinates and is not enough for actually operating a manipulator to hold the part. For actual handling, both the position to be held and the direction which the manipulator approach to the position in should be determined three-dimensionally.

In this paper, we propose an extended method of obtaining the three-dimensional information required to hold clothes by the second manipulator, when the clothes is held in the air at any point by the first manipulator. The biggest contribution of this paper is calculation of the appropriate grip coordinates for holding a target part based on the predicted 3D clothes shape. After explaining the algorithm in Section II∼VI, we show some results of preliminary experiments using actual two manipulators.

II. Flow of vision module

We suppose a system consisting of two manipulators and two cameras which are calibrated relative to the manipulators. Figure 1 shows an example of such system, which we actually used for our experiments. The task we consider in this paper is to hold up a target part of clothes (e.g. one shoulder of a pullover) by the second manipulator, under the situation that the clothes is held in the air at any point close

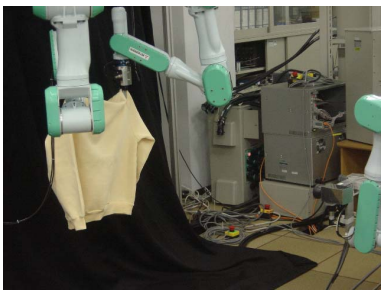

Figure 1: System consisting two manipulators and stereo cameras

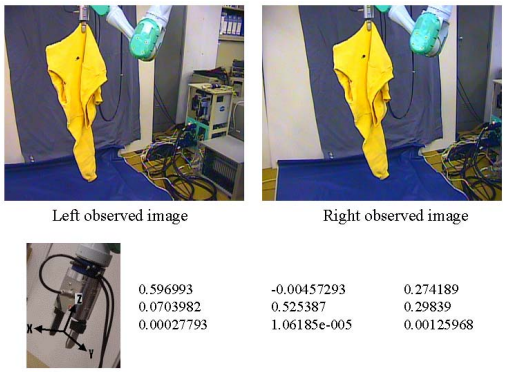

Figure 2: Example of input to Vision module

to its hem by one manipulator. Although the task is simple, by iterating the task a few times, the system can hold the clothes at a goal state (e.g. opening a pullover by holding it at its two shoulders)[4]. We assume that we know style of the target clothes (e.g. pullover, trousers) and its approximate sizes in advance.

The input to the vision module is left and right images observed by stereo cameras and the position and pose of the first manipulator which holds the clothes. Concretely, the latter is given in the term of the grip coordinates in the world coordinates as shown in Fig. . The output of the module is the grip coordinates for holding action with the second manipulator.

The flow of the vision module for the task is as follows:

1. Prediction of possible 3D shapes of the clothes
2. Estimation of the clothes states using the left and right images respectively
3. Calculation of the grip coordinates of the second manipulator for holding-up action

—-Manipulator action————

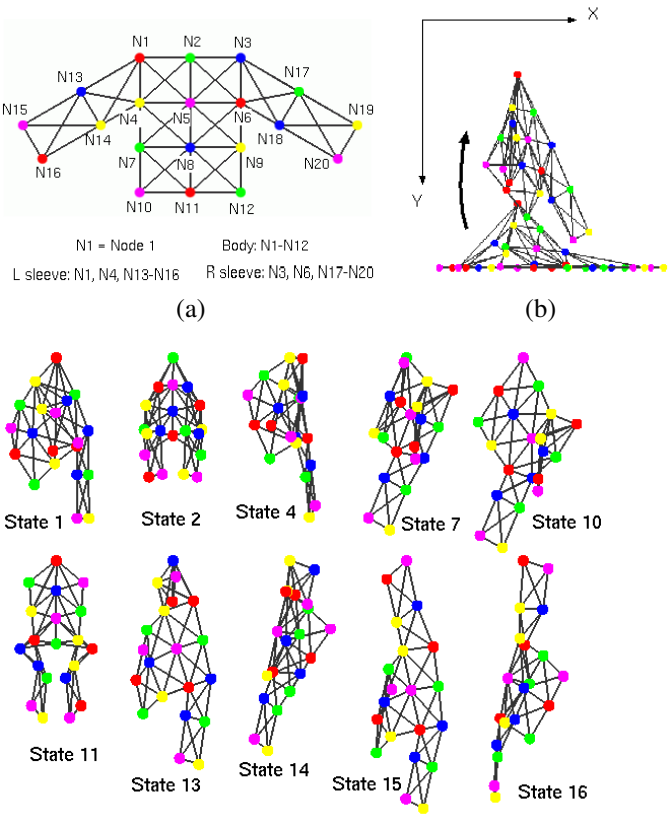

Figure 3: Model shapes: (a) common pullover model; (b) simulation processes; (c) possible 3D shapes when the pullover is held at a point

4. Verification of the current estimation by comparing new observation with prediction

Each process is explained in the following sections.

III. Prediction of possible 3D shapes

To simulate the clothes deformation only from the knowledge of its style and approximate sizes, a common model for each style is stored in advance. Fig. 3a shows a common pullover model. To make the problem simple, currently, we assume the front and back sides of the clothes are not separated and no thickness is given to the model. The model consists of 20 nodes which are connected to each other by springs as illustrated with the lines. The details on the spring setting can be seen in [4]. An individual deformable model for the target clothes is automatically built based on its approximate sizes, for example,like the width and length of the trunk and the length of the sleeves for a pullover.

Possible deformed shapes when the pullover is held in the air at a point close to the hem is automatically obtained

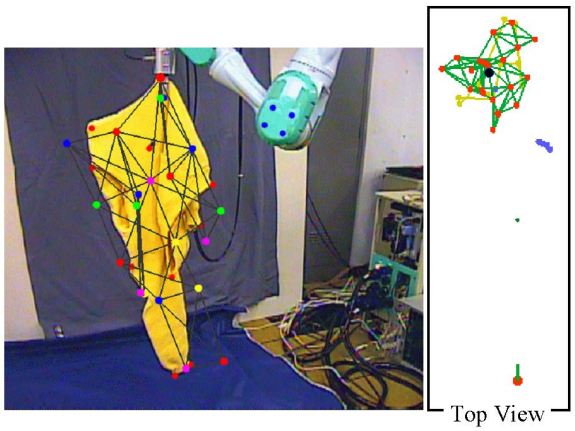

Figure 4: Example of 3D predicted shape set at the holding position and its projection to an observed image

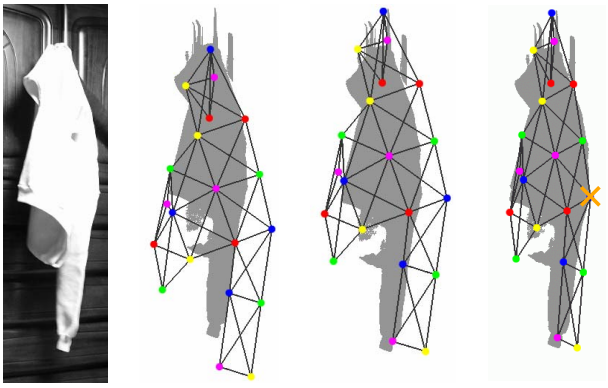

Figure 5: model adjustment: (a) observed image; (b) original position (gray region shows observed clothes region); (c) after vertical translation; (d) after width normalization.

through the following simulation. First, the model is virtually spread on a horizontal plane, under the situation that the gravitational forces are constantly exerted all nodes. The deformation of the pullover is simulated while a node of the model is moved up vertically little by little (Fig. 3b). Holding at just a point produces uncertainty regarding the rotation around the gravity direction. The actual holding we assume is not by a point but by a flat grip with a small area, so the direction of the grip specifies the rotation. By assuming the pullover is observed from the direction perpendicular to the grip plane, the viewing direction is fixed so that it coincides with the normal of the plane around the holding point. Fig. 3c shows examples of the predicted 3D shapes, from State 1 to State 20, named after the number of the node held. State 3, 6, 9, 12, 17, 18, 19 and 20 are left-right symmetrical shapes of State 1, 4, 7, 10, 13, 14, 15 and 16, respectively.

IV. Estimation of clothes states

First, each possible 3D shape as shown in Fig. 3c is set at the holding position which is given by the input grip coordinates of the first manipulator. Figure 4 shows a top view when one possible 3D shape is placed at the holding position. In the top view, the red circle at the lower part of the figure and the green line sticking out from the circle illustrate the position and the view direction of the stereo camera system; the black circle illustrates the grip position of the first manipulator; red points connected by green lines show the predicted shape of the hanging clothes. Then, the predicted appearances in the left and right images are calculated by projecting the 3D shape at the position to the images. In the left image of Fig. 4, an example of such predicted appearances is superimposed in an observed image.

Next, the following processes are applied to the left and right images respectively using each predicted appearance as a model appearance. Here, we assume the clothes regions in observed images can be extracted in a bottom-up way. The gray region in Fig. 5b shows an example in the case of the observed image in Fig. 5a.

i) Selection using simple attributes
By checking attributes of the appearance which can be robustly extracted from observed images, model appearances which are clearly different from the observed appearance are excluded from the candidates. At the current, the vertical length of appearances is the only one attribute we use for this process.

ii) Adjustment of model appearance
In order to compensate the difference between the actual and the predicted appearances, the appearance is adjusted to the observed region as follows:

(a) Vertical translation
The observed clothes may be held at a point far from any 20 nodes which are used as holding positions to predict representative model shapes. In the case, the observed appearance is similar to the appearance of the representative shape of one of the closest node, but is misaligned along the vertical direction. To remove such misalignment, model appearances having longer vertical length than the observed one are vertically translated up on the image by difference in their lengths. Fig. 5c shows the model appearance after this translation.

(b) Width normalization
Each model appearance is horizontally shrunk or extended so as to have the same width as the observed region. Here, the shrinkage/extension ratios are separately calculated for the left and right parts relative to the holding position. Fig. 5d shows a result after this normalization.

iii) State decision using overlap ratio
After these adjustments, consistency between the observed and model appearances is checked. Currently the overlap ratio, R, which is the sum of the ratio of overlapped area to model appearance area and the ratio of overlapped area to observed area, is used. The model state which has the highest value of R is selected as the estimation result.

iv) Detection of the image coordinates of the point to be held next
The position corresponding to the part to be held next on the observed image is detected by searching the edge point closest to the model node corresponding to the part. In the pullover case, the edge point closest to one of the shoulder nodes is indicated if it is easy to grasp, that is, it is non-concaved and non-overlapped in the model appearance[4]. The orange cross in Fig. 5d shows the detected position in this example.

During this detection process, the consistency of the contour shape around the point is checked to improve the robustness. If inconsistency is detected, the estimation is excluded. Then, the model appearance having the next highest value of R is reselected and tried on the process iv).

The estimation results from the left and right images are synthesized to give a final judgment.

V. Calculation of grip coordinates for grasping action

If the clothes states are consistently estimated in the left and right images, next, the grip coordinates for grasping the indicated point are calculated according to the estimated model shape. The grip coordinates of our system is shown in Fig. . To grasp the point, the second manipulator approaches along the minus Z direction to the part with the Y axis set to coincide with the normal of the part. When the origin of the coordinates reaches to the position to grasp, the grip is closed. The grip coordinates at the grasping position, which satisfy these conditions can be calculated as follows:

i) Calculation of the 3D coordinates of the indicated point
Although the target points are indicated both in the left and right images, they do not necessarily correspond to each other exactly. To obtain more accurate stereo point pairs, the edge point in the right image is searched along the epipolar line corresponding to the image coordinates of the target point in the left image. Then the 3D coordinates of the corresponding point are calculated based on stereo matching principle.

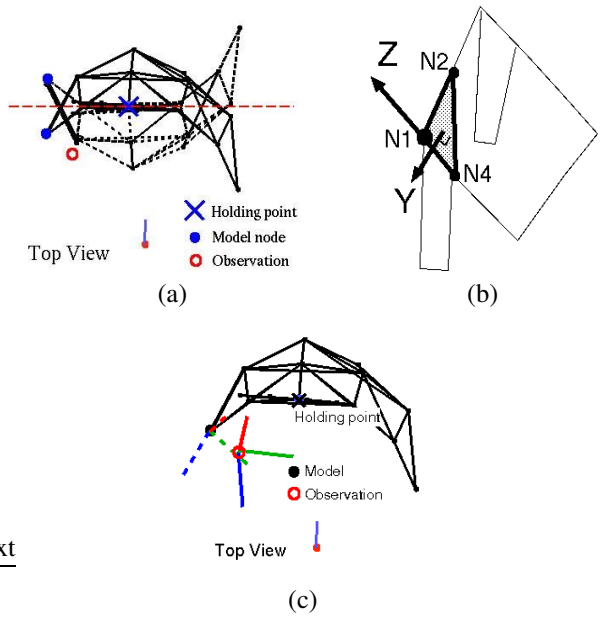

Figure 6: Determination of grip coordinates: (a) Decision of model shape from symmetrical candidates; (b) calculation of the coordinates based on model shape; (c) correction based on the observation.

ii) Decision of model shape from symmetrical candidates
Notice that, if we assume weak-perspective, for one appearance, there are two possible 3D shapes which are symmetrical with respect to the vertical plane which are perpendicular to the view direction and goes through the current holding point. Figure 6a is a top view illustrating the two symmetrical model shapes: one is shown by solid lines, while the other is by dotted lines. The cross shows the current holding position. From the process i), the actual 3D position corresponding to the model nodes shown by blue dots was calculated. The model shape which has closer model node to the observation is selected. For example, if the actual position is lied as shown by the red circle in Fig. 6a, the shape shown by the solid lines is selected.

iii) Calculation of grip coordinates based on predicted shape
In order to determine the direction of the grip coordinates, the posture of the part to be held should be known. Although the surface normal of the part is required for that, it is not easy to obtain the normal direction of free-form surface from stereo camera systems. Instead of such bottom-up way, we propose

to estimate the posture of the part from the predicted 3D shape in a model-driven way as follows. Fig. 6b shows an example when the point to be held is one shoulder of a pullover. The other shoulder, Node 1, is a target point to be held next. The normal direction of the adjacent triangle patch, that is the triangle surrounded by Node 1, 2 and 4 in this case, is used for the direction of the Y axis. The direction of the Z axis, which is approaching direction, is set to coincide with the line connecting Node 1 and 4. The X axis is determined to complete the right-handed coordinate system. The red, green and blue dashed lines in Fig. 6c show respectively the X, Y and Z directions obtained in this example.

iv) Correction of grip coordinates based on observation
The actual clothes shape may be different from the model shape. Therefore, if any good clues indicating the actual posture can be found in the observed images, the grip coordinates is corrected based on the information. Currently, the 3D line corresponding to the model contour adjacent to the part to hold, which is the line connecting Node 1 and 2 in the case of Fig. 6b, is obtained based on the stereo matching of the left and right images. Suppose that the 3D line is obtained as shown with the red line in Fig. 6c. Then, the first estimated grip coordinates are rotated to fit the observation as shown with the red, green and blue solid lines in Fig. 6c.

After the direction of the grip coordinates is fixed, its origin is set a few centimeters inside along the Z direction from the 3D position of the edge point.

v) Calculation of goal grip coordinates
As the goal position after grasping, we use a grip coordinates which is parallel to the grip coordinates of the first manipulator and far from it in the horizontal direction vertical to the view direction. The distance from the first grip is set the same as the distance between the 3D position to the target point and the position held by the first manipulator.

VI. Verification

The appearance of the pullover after the action by the second manipulator is simulated using the same deformable clothes model. The method can judge if the current estimation of the state is correct or not by comparing the expected appearance with the newly observed appearance.

VII. Experimental results

A. Experiments on estimation

First, for the purpose of examining the ratio of correct estimation, we have conducted experiments using single images of a pullover. which were held in the air at a point by a human hand. In the experiments, the largest uniform color region under the holding position was extracted as the observed clothes region.

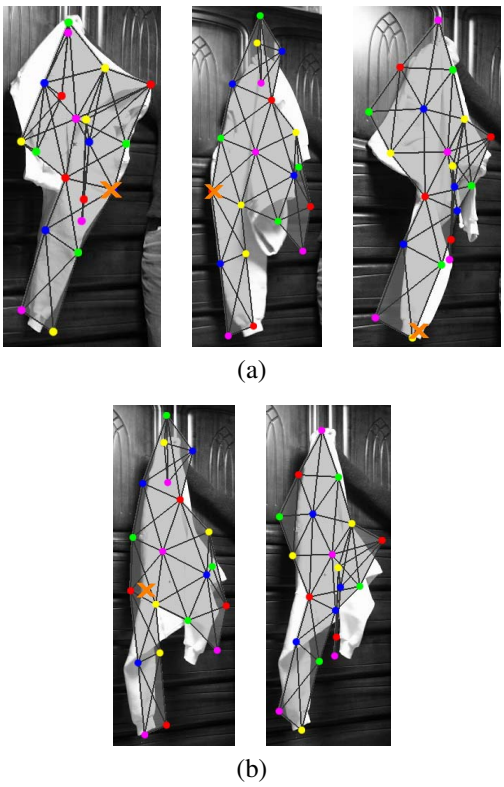

Figure 7: Estimation results on a pullover: (a) examples of success; (b) example of failure (The left one is selected, while the right one is correct).

We used 25 images which were taken while holding the pullover at the position corresponding to any nodes or any middle points between two adjacent nodes of the model. For 19 images, the correct state was selected with the first priority after the overlap ratio check (process iii) in Section IV). For the remaining six images, the correct state was selected with the second or third priority. Since the consistency check around the point to be held next detected the wrong states in three of the six images, at the final, for 22 images, the correct state was selected. Fig. 7a shows three examples of success. The point indicated for the next grasp action is marked with the orange cross. In the most right case, the lowest point of the observed region was indicated since the shoulder of the estimated state is hard to grasp. In the remaining three images, the state was wrongly estimated. Fig. 7b shows one example of failure. The left and right images of Fig. 7b show the selected and the correct state respectively. The correct state was selected as the second.

B. Experiments using manipulators

We conducted preliminary experiments using an actual

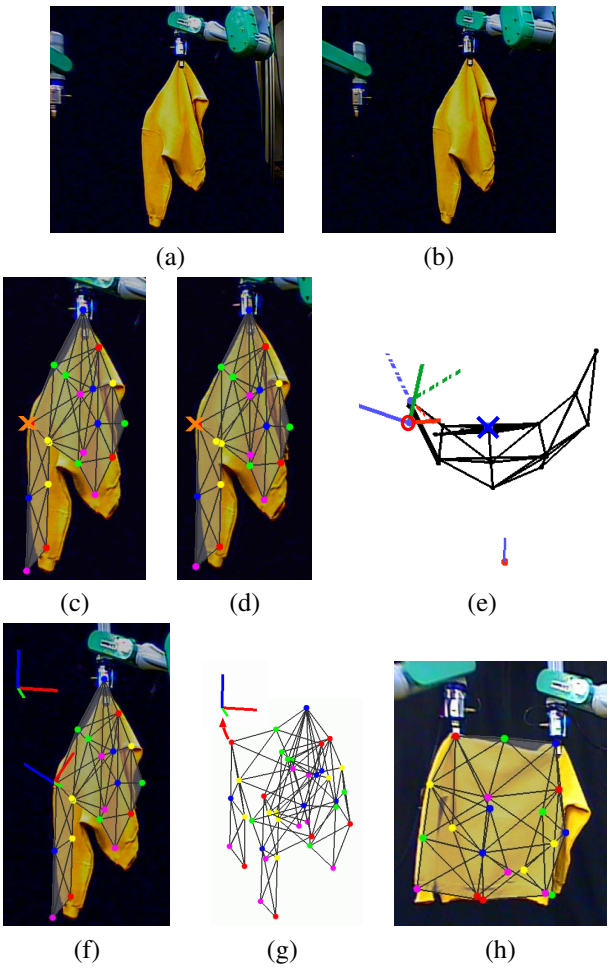

Figure 8: Handling experiment using relatively hard pullover: (a) observed left image; (b) observed right image; (c) estimation result(left); (d) estimation result(right); (e) resultant grip coordinates in a top view; (f) grasp and goal grip coordinates in the left image; (g) simulation on the action; (h) newly observed image with the predicted appearance superimposed.

system consisting of stereo cameras and two manipulators (Mitsubishi PA-10).

First, we use the same pullover which was used in the experiments described in the previous section. Fig. 8a,b show observed left and right images. Fig. 8c,d show the estimation results of the both images. The selected state was State 3. The points to be held next were indicated as marked with the orange crosses. The 3D coordinates corresponding to the indicated points were calculated as shown with the red circle in the top view (Fig. 8e). Predicted 3D shape selected from two symmetrical candidate shapes is shown with solid lines in Fig. 8e. The grip coordinates for next action which were estimated from the predicted shape were shown in Fig. 8e,f: the red, green and blue lines correspond to the X, Y and Z axis of the grip coordinates respectively: the dashed and solid lines correspond to the coordinates calculated from the model shape and the corrected one using the observed shoulder line respectively. When conducting this experiment, we have not yet implemented the correction of the grip coordinates using the observation. Therefore, the manipulator approached to the shoulder edge shown by the red circle from the direction parallel to the dashed blue line instead of the solid blue line. Although the direction was about 40 degrees different from the actual tangent direction of the part, flexibility of the clothes absorbed the error in the approaching direction and the part was successfully grasped. From this observation, it is expected the allowable error in the approaching direction can be fairly wide from the practical view point. However, we cannot say anything about the amount of allowability because of the low number of the observations up to current times. The difference in the Z axis of the corrected grip coordinates and the actual tangent direction is less than 5 degrees. Therefore, we are sure that the grasping should have also succeeded if we adopted the corrected values. Fig. 8g shows the simulation processes for predicting the clothes shape after the action executed. Fig. 8h shows the newly observed images after the action with the predicted appearance overlapped on it. The good coincidence shows the correctness of the estimation on the clothes state.

Fig. 9 shows experimental results using a toddler pullover which is softer than the previous one. The individual pullover model is automatically built based on the width and length of the trunk and the length of the sleeves. Fig. 9a,b show observed left and right images. Fig. 9c,d show the estimation results of the both images, which were State 17. Notice that the model appearances were largely shrunk horizontally to fit the observed appearances, since the actual 3D shape was folded more than the model shape. (You can see the original width of the model appearance in Fig. 9g.) Fig. 9e,f show the grip coordinates for next action which were calculated based on the predicted 3D shape. Because the large folding actually happened, the 3D position corresponding to the shoulder node was far from the position of the model shape. In this case, since the shoulder line was almost vertical and did not give any clues on the tangent direction of the part, the direction of the grip coordinates determined from the model shape was used as it was. Although the actual tangent direction of the shoulder part was different from the approaching direction by about 20 degrees, the difference did not disturb the grasping action. As a result, the part was successfully grasped. Fig. 9h shows the newly observed image after the action with the predicted appearance, which was obtained through the simulation shown in Fig. 9g.

VIII. Conclusions

We proposed a deformable model-driven method to ob-

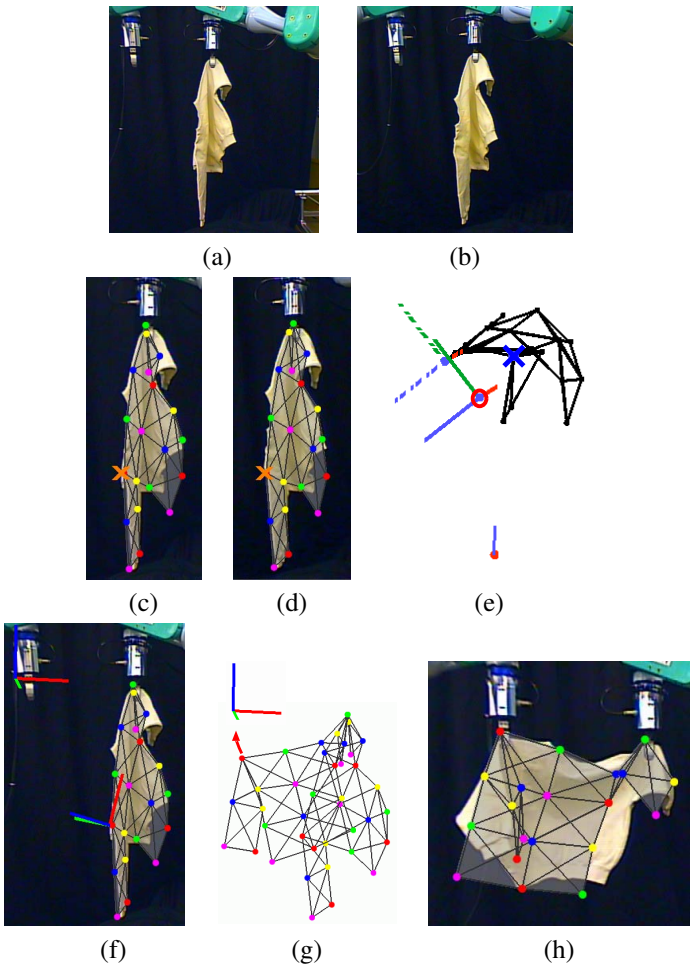

Figure 9: Handling experiment using relatively soft pullover: (a) observed left image; (b) observed right image; (c) estimation result(left); (d) estimation result(right); (e) resultant grip coordinates in a top view; (f) grasp and goal grip coordinates in the left image; (g) simulation on the action; (h) newly observed image with the predicted appearance superimposed.

tain the 3D information necessary for grasping a target part of clothes by manipulators from observation with stereo cameras. The method enables the estimation of the state of largely deformed clothes held at a point by predicting possible 3D shapes through the simulation with a deformable model. The strategy also has the following good aspects:
1) The robustness of handling processes can be improved by verification processes through the simulation using the model.
2) The 3D information which is required for handling action but hard to obtain from images in a bottom-up way can be estimated from the predicted 3D shape.
The experiments using actual manipulators have shown the good potential of the proposed method. More experiments is necessary to clarify its ability and limitations.

In the experiments of this paper, we use simple elastic models for predicting the deformation of the clothes. As a result, the predicted shapes are fairly rough. Although the strategy works well with the rough predicted shapes, we believe that the closer prediction should improve the ratio of correct estimation in the same scheme. The accuracy of the predicted shape can be improved by using more sophisticated clothes models, although we should consider the trade-off between the computational complexity and the prediction accuracy. In addition, simple models have a good point that they require less prior knowledge of the clothes in question.

Our future work includes more collaboration with manipulations. For example, "shaking by a manipulator" can be effective to remove unstable deformations. We also plan to use manipulator action to extend the clothes when it is folded too much and its appearance hardly gives any clues on the state.

Acknowledgment

We are thankful to Dr. K. Tanie, Dr. K. Sakaue, and Dr. T. Suehiro for their support to this research.

References

[1] M. Inaba and H. Inoue: "Hand eye coordination in rope handling", *JRSJ (Journal of Robotics Society of Japan)*, Vol. 3, No. 6, pp. 32–41, 1985.

[2] D. H. House and D. E. Breen: *"Cloth modeling and animation"*, A. K. Peters, Ltd., 2000.

[3] M. Kaneko and M. Kakikura: "Plannning 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.

[4] Y. Kita and N. Kita:"A model-driven method of estimating the state of clothes for manipulating it", In *Proc. of 6th Workshop on Applications of Computer Vision*, pp.63–69, 2002.

[5] M. Yamamoto, P. Boulanger and et al.: "Direct estimation of deformable motion parameters from range image sequence", In *Proc. of 3rd International Conference on Computer Vision*, pp. 460–464, 1990.