3D Shape Measurement of a Large Cloth Extremely Close to a Fisheye Stereo Camera

Nobuyuki Kita

Abstract—Existing 3D stereo shape measurement techniques do not consider the case of a large item very close to the stereo camera (i.e., covering almost the entire field of view). This case is crucial for humanoid robots with a head-mounted stereo camera performing support tasks because this is a natural configuration when handling large objects (e.g., adult clothes, bath towels, and linens). To address this, this paper describes the development of a 3D shape measurement method using a fisheye stereo setup. The method is successfully applied to measure the 3D shape of a large coat hung over the arms of a humanoid robot. However, as the object approaches the camera, new 3D shape measurement challenges arise. The causes of these difficulties are clarified, and suitable improvements to the algorithms are proposed. Experiments validate the effectiveness of the proposed improvements.

I. INTRODUCTION

In the near future, robots will be expected to perform daily household support tasks as the number of older adults increases. Measuring the 3D shape information of objects is indispensable in enabling humanoid robots to handle large objects. The handling of large objects is required to execute tasks such as dressing activities, managing laundry, and moving furniture. A common and challenging case involves the handled object being very close to the eyes of the robot, covering almost the entire field of view. Many stereo measurement methods have been developed and published, but no existing methods consider an object covering almost the entire field of view. Additionally, when an object is extremely close to the imaging plane, its appearance may vary vastly between the right and left images (Fig. 1). The author has developed a method to measure the 3D shape of an object from stereo images captured by a fisheye stereo camera mounted on the head of a humanoid robot [1][2]. To cope with the substantial differences of the object’s appearance between the right and left images, the local appearance in the right and left images is first predicted based on the local appearance in the left and right images, respectively. The local-area dissimilarities are calculated along an epipolar line using the predicted appearance, and the correspondence matches are searched. The method obtains suitable results when an object is 0.4 m away from the stereo camera whose baseline length is 0.15 m.

Depending on the size of the object, the length of the robot’s arms, and the purpose of the object handling, the humanoid robot may be required to handle an object much closer to its head. When the above method is applied to fisheye stereo images of an object that is 0.2 m away from the stereo camera whose baseline length is 0.15 m, new stereo measurement challenges arise. The causes of these difficulties are clarified in this paper and improvements to the algorithms are proposed accordingly. Close measurement and extremely close measurement refer to the measurement of an object that is 0.4 m and 0.2 m away from the stereo camera, respectively. The utility and effectiveness of the proposed improvements are verified experimentally.

Section 2 explains the method for close measurement. The causes of challenges arising for extremely close measurement are investigated in Section 3. Algorithm improvements are proposed in Section 4. Section 5 shows the experimental results and evaluations. Section 6 provides a summary and conclusion.

II. METHOD FOR CLOSE MEASUREMENT

This section provides an overview and brief description of the methodology of close measurement. For more detail, Section 2.1 describes its principles [1] and Section 2.2 briefly explains the procedure of the method which is applied to a cloth hung by a humanoid [2].
A. Overview

All of the intrinsic and extrinsic parameters of the fisheye stereo camera mounted on the head of a humanoid robot are calibrated in advance. The stereo correspondences are searched along the epipolar lines. The local appearance of the object is substantially different between the right and left images. However, the object appearance along an epipolar line in one image (subimage) can be predicted from the appearance in the other image (main image) using homography if the local area can be approximated as a plane and the surface normal of the local plane is known (Fig. 2). The method for close measurement is based on block-matching between the actual appearances in the subimage and the predicted appearance from the main image.

B. Close Measurement of a Cloth

Here, the algorithm for close measurement of a large cloth hung in the air by one or both hands of the humanoid robot is briefly explained. More details are provided in Section 4.

1) Possible area and fundamental normal direction (Fig. 3): The approximate size of the cloth is assumed to be known. The real-world 3D position of the cloth can be obtained from the positions of the robot’s hands. This 3D area is henceforth referred to as the possible area. The fundamental normal direction of the cloth can be determined from the position relationship between the two hands when the cloth is hung over both hands or from the pose of the hand when the cloth is hung over one hand.

2) Setup of matching candidates: Matching candidates are set in a grid pattern with intervals ranging between one and a few pixels within the projected 2D area of the possible area in the main image (Fig. 5(a)). Each matching candidate is linked to the four neighboring matching candidates.

3) Setup of depth candidates: For each matching candidate, the depth candidates are set with equal depth intervals on a 3D line segment that is the part of the light ray corresponding to the matching candidate within the possible area (Fig. 6(a)).

4) Generation of predicted block: For each depth candidate, the appearance of the block whose center is its projected position in the subimage is predicted from the appearance in the main image by using the homography that can be obtained from the fundamental normal direction and the 3D position of the depth candidate (Fig. 14). The block that is predicted is called the predicted block and the part of a main image that is used for the prediction is called the source area in the remainder of this paper.

5) Calculation of dissimilarity: A depth candidate is projected at a real coordinate in the subimage. For each depth candidate, the normalized sums of squared differences (NSSDs) are calculated between the predicted block and the four blocks whose central pixels are near to the real coordinates. The dissimilarity of the depth candidate is obtained through bilinear interpolation of the four NSSDs.

6) Determining the f-match: For each matching candidate, the sequence of dissimilarities of the depth candidates is verified and determined to be an f-match when the sequence has only one clear peak satisfying a threshold. The depth for the f-match is fixed at the depth of the depth candidate corresponding to the peak. When the sequence has more than one peak satisfying a threshold, the matching candidate is categorized into multiple peaks.

7) Determining the v-match: Since the cloth is pulled downward by the gravity, vertical neighborhoods are expected to have the similar horizontal position. This is called vertical constraint in the reminder of this paper. A matching candidate in the set of multiple peaks is determined to be a v-match if it is consistent with respect to the vertical constraint.

8) Determining the h-match: In the horizontal direction, there are gentle concavity and convexity like drape. The horizontal position and the surface normal direction of horizontal neighborhoods are expected to vary gently and continuously. This is called horizontal constraint in the reminder of this paper. For each matching candidate that is neither an f-match nor a v-match, the predicted blocks for all depth candidates are regenerated using a new normal direction that is induced from the horizontal constraint; the dissimilarities are then recalculated. If the sequence of dissimilarities has only one clear peak satisfying a threshold, the matching candidate is determined to be an h-match.

9) Output: The output comprises the 3D positions of the f-match, v-match, and h-match.
III. DIFFICULTIES ARISING FROM EXTREMELY CLOSE MEASUREMENT

Fig. 4a and b depict the positional relationship between the cameras (black rectangles) and cloth (thick brown lines) for extremely close measurement and close measurement, respectively. The black lines emanating from the right camera represent the light rays projected on the right image at equal intervals. The light rectangles signify the possible areas. The following new challenges for extremely close measurement are demonstrated in these figures.

- The projection of the cloth expands to cover the periphery of the images.
- The areas that are visible from one camera but invisible from the other camera (thick yellow and green lines) increase.
- The angle included in the light ray search range (thick blue line) becomes large.
- Within the area surrounding the small yellow circle, the distances from the left and right cameras are quite different and the cloth is projected onto the center of one image and the periphery of the other image.

IV. ALGORITHM IMPROVEMENTS

The necessary improvements to the algorithms addressing the difficulties described in Section 3 are explained here. For convenience, the right image is the main image and the left image is the subimage in the discussion.

A. Setup of Matching Candidates

For close measurement, matching candidates are set in a grid pattern with intervals between one and a few pixels on the main image (Fig. 5a (left)). Each matching candidate is linked to the four neighboring matching candidates. Owing to the first challenge enumerated in Section 3, some matching candidates are set in the extreme periphery. Although they correspond to equal intervals in the main image, their intervals in the 3D space vary substantially because of strong radial distortion as shown in the right subimage of Fig. 5a. Furthermore, the relative positions between the four neighboring matching candidates deviate from the vertical or horizontal relationship.

In the improved algorithm, the matching candidates are set in a grid pattern with given intervals over the center of the plane of the possible area (see the left image of Fig. 5b, the interval is enlarged in the image for clarity but is much smaller in reality). They are then projected onto the main image. The matching candidates are set at the pixels corresponding to the projected positions. If more than one projected position corresponds to the same pixel, only one matching candidate is generated.

B. Setup of Depth Candidates

Fig. 6 depicts the approach to set the depth candidates (red dots) for each matching candidate. For close measurement, the candidates are set at equal depth intervals on a light ray (blue line in Fig. 6a) corresponding to the matching candidate (in the figure, the interval is enlarged for clarity but is much smaller in reality). Owing to the third challenge listed in Section 3, the intervals of the projected positions in the subimage vary substantially, as shown at the bottom of Fig. 6a. The improved algorithm sets depth candidates such that their projection on the subimage yields equal intervals, as shown in Fig. 6b (the interval is enlarged in the figure for clarity but is one pixel in reality).
C. Generation of Predicted Block

For each depth candidate, the close measurement method always generates the predicted block from the source area of the main image. In Fig. 7, the short thick red line segments represent the 3D locations corresponding to the backward projection of the predicted block of some depth candidates on a light ray (dashed blue line) corresponding to a matching candidate. Since the short red lines are surrounded by the black lines emanating from the left camera, their projections on the left image have the same size. However, the sizes of their projections on the right image are vastly different as demonstrated upon a comparison of the short red lines and the gray lines emanating from the right camera. In particular, owing to the fourth challenge listed in Section 3, the projection on the right image becomes very small when the position of the depth candidate is just in front of the left camera. That is, the source area for the predicted block of such a depth candidate is very small and the image conversion from the source area to the predicted block involves a substantial enlargement. In this case, the amount of information in the predicted block is reduced and the quality of the dissimilarity calculation decreases. The improved algorithm does not fix the direction of the prediction and changes it so that the image conversion from the source area to the predicted block could be always shrinking (Fig. 15). The close measurement method uses the bilinear method to shrink the image source. When substantial shrinking is involved, bilinear image-shrinking loses source area information. Instead, the improved algorithm uses the area-average method (Fig. 16).

D. Calculation of Dissimilarity

The number of depth candidates is proportional to the length of the epipolar line following the improvement described in Section 4.B. The matching candidates having a long epipolar lines have many depth candidates (as a result of the first challenge listed in Section 3). The number of matching candidates that are invisible in the subimage increases as a result of the second challenge listed in Section 3. It is thus necessary to improve the dissimilarity calculation.

When the observation target is close to the stereo camera, the appearance of the target shows brightness differences between the right and left images because the relationship between the target, the right/left camera, and the lighting system differ. NSSD is robust to these unfavorable differences. However, this robustness comes at the cost of large differences in average brightness and brightness variations resulting in incorrect matches. To avoid such incorrect matches, the improved algorithm adds a penalty to the value calculated by NSSD when the difference between the average brightness or the brightness variation exceeds a threshold.

The calibration of all of the intrinsic and extrinsic parameters is very carefully conducted through a re-projection error minimization using a calibration target. However, there is still a re-projection error of approximately one pixel on average and two pixels at the maximum. The correct position may therefore deviate by one or two pixels from the epipolar line. The improved algorithm calculates the dissimilarities between the predicted block and the source block not only on the epipolar line but also at four additional positions vertically deviating by one or two pixels from the epipolar line. The minimum value is set as the dissimilarity for the depth candidate.

E. Determining the F-match

A matching candidate is judged as an f-match when the sequence of dissimilarities of the depth candidates has only one clear peak satisfying the threshold. To judge the strength of the peak, the differences between the peak value and the values of two neighboring depth candidates are evaluated. Prior to the improvement described in Section 4.B, any depth candidate has constant depth differences with the neighboring depth candidates, and the threshold is constant. Following the
improvement in Section 4.B, any depth candidate has a one-pixel difference between the neighboring depth candidates in the subimage. However, when the predicted block is generated from a source area in the subimage through a $1/s$ shrinking conversion (where $s$ is the scale factor), the interval distance in the subimage becomes $1/s$ pixels. The improved algorithm evaluates the differences between the peak value and the values of two depth candidates that are the $s$-pixel neighbors.

F. Determining the V-match

Each matching candidate in the set of multiple peaks is determined to be a v-match candidate if one of its vertical neighbors is an f-match. The horizontal distance between vertical neighbors is small according to the vertical constraint. The depth candidate whose dissimilarity value is lower than the threshold and whose 3D position is closest to the 3D position of the f-match is judged as the v-match if the horizontal distance between those 3D positions is lower than the threshold. The improved algorithm determines the v-match as the depth candidate whose dissimilarity is the minimum and is lower than the threshold among the candidates whose horizontal distance from the 3D position of the f-match is lower than the threshold.

G. Determining the H-match

Each matching candidate that is neither an f-match nor a v-match is an h-match candidate when one of its horizontal neighbors has a fixed 3D position that is an f-match or a v-match. Horizontal neighbors are continuous according to the horizontal constraint. For each depth candidate, the predicted block is regenerated by using a new surface normal direction recalculated by using the 3D position of a horizontal neighbor. Owing to the improvement in Section 4.B, the 3D intervals between the depth candidates are not constant and increase at the periphery. In this case, the surface normal directions (red arrows) that are recalculated by using the 3D position of a horizontal neighbor (red dot) change substantially between neighboring depth candidates, as shown by the green arrow in Fig. 8. The improved algorithm first reproduces the depth candidates such that their surface normal changes constantly between the neighboring depth candidates, as shown on the right of Fig. 8, then generates the predicted block and calculates the dissimilarities for each depth candidate. A matching candidate is determined to be an f-match when the new sequence of dissimilarities of depth candidates has only one clear peak satisfying the threshold.

V. EXPERIMENTS

A. Evaluation of the Suite of Algorithm Improvements

To evaluate the challenges of extremely close measurement and the effect of the improvements proposed in Section 4, two sets of stereo images are used (Fig. 9). One set depicts a cloth 0.4 m away from the stereo camera and the other depicts a cloth 0.2 m away. The shape of the cloth is changed as little as possible between the two positions. The following three experiments are performed.

a. The algorithm with only the improvement described in Section 4.D is used to process the input images at the 0.4 m distance.

b. The algorithm with only the improvement described in Section 4.D is used to process the input images at the 0.2 m distance.

c. The algorithm with all of the improvements is used to process the input images at the 0.2 m distance.

Figs. 10, 11, and 12 show the results of experiments a, b, and c, respectively. The top row of each figure shows the matching results where yellow, blue, and red denote f-match, v-match, and h-match, respectively. The middle row of each figure shows the 3D measurement results where the intensity changes correspond to the depth changes with respect to the stereo camera. The bottom row of each figure shows the results of texture mapping where only the square patches with fixed depths at their four corners are mapped. Table 1 shows the total number of matching candidates and the percentages of f-match, v-match, h-match, and no match.

Experiment a achieves approximately 70% correspondence with matching candidates. Since there do not appear to be many incorrect matches in the 3D views, most of the correspondences appear to be correct and the result looks adequate. In contrast, the percentage of matching candidates
with correspondences decreases to approximately 48% for experiment b, and many of the correspondences appear to be incorrect. The quality of the 3D views and texture mapping are severely deteriorated. In experiment c, the percentage of matching candidates with correspondences increases to approximately 79%, with the biggest increase attributed to the v-match and h-match percentages. Incorrect matches appear to be decreased. The quality of the 3D views and texture mapping is recovered.

The 3D shape measurement results are horizontally sliced into 50-mm sections and shown in Fig. 13. The left, middle, and right columns show the results of a, b, and c, respectively. The result of experiment a is more reliable than the others since the image-capture conditions are better than for the others (i.e., the cloth is farther from the camera). The results of experiments b and c are compared with the result of experiment a. Compared with a, experiment b resulted in only limited 3D depth information, especially in the periphery. There are many outliers because of incorrect matching. There are also many 3D depths with small errors caused by the low quality of the matching. In contrast, experiment c resulted in correct 3D depths at almost the same parts as were obtained in experiment a, including the periphery, and the missing parts were mostly caused by increasing occlusion. The number of outliers increased compared with experiment a, but the number of 3D depths with small errors is drastically decreased compared with experiment b.

B. Evaluation of the Improvements in Section 4.C

Here, we show the results of different dissimilarity calculation approaches after generating the predicted blocks (11×11 pixels) for every depth candidate for a particular matching candidate. Fig. 14 shows the results before applying the improvements discussed in Section 4.C. The grey-level images with the light-blue background correspond to the right image and those with the light-pink background correspond to the left image. The predicted blocks are always generated from the source areas in the right image and the dissimilarities are calculated between the predicted blocks and the source blocks in the left image. The chart in Fig. 14 shows the dissimilarities. The horizontal axis represents the depth candidates which are numbered in order of increasing depth (i.e., the closest depth candidate is labeled with “1”). The vertical axis represents the dissimilarity values. Since the dissimilarity threshold is 0.5, several depth candidates are potential matches; therefore, the matching candidate is categorized as multiple peaks. The source areas and the predicted blocks are depicted for the eight depth candidates whose dissimilarities are small. The yellow rectangles in the top row of Fig. 14 show the position of the source blocks/areas of the eight depth candidates. The source blocks/areas are enlarged and shown in the lower row. The predicted blocks are also enlarged with the same scale and shown in the lower row. The predicted blocks are generated from the source areas through image enlargement with the exception of the first one.

Fig. 15 shows the results following the improvements...
described in Section 4.C for the same scenario as is shown in Fig. 14. Here, the enlargement scale of the source areas in the left image is half that of the others. The predicted blocks are generated from the source areas through image shrinking on the left image except for the first depth candidate. After the improvements, the dissimilarities for the six depth candidates whose dissimilarities were very small before the improvement become large and the number of depth candidates with dissimilarities below 0.5 decreased from eight to two.

Fig. 16 shows the predicted blocks in the left image and the chart plotting the dissimilarity values. The predicted blocks are generated from the same source areas as the former experiment by using the area-average method. It is difficult to see the predicted block differences between Figs. 15 and 16, but the dissimilarity for the first depth candidate increases and the number of depth candidates with dissimilarities below the threshold becomes just one.

VI. SUMMARY

This paper presents an approach to improve 3D shape measurement using a fisheye stereo camera setup to enable the shape measurement of a very close object. This is particularly essential for humanoid robots performing household tasks. The utility of the method is verified in actual experiments. Although some incorrect matches remain for a cloth that is 0.2 m away from a stereo camera, these can be alleviated with some straightforward post-processing. Comparative analysis of the results reveals that the suggested improvements significantly improve the amount of accurate information obtained by the measurement technique. Whereas the conventional methodology yields unusable results at 0.2 m from the camera, the improvements allow for a reconstruction that is comparable to that achieved for a 0.4-m distance.

Reducing the calculation time is one important topic to widen the applicability of this methodology. One plausible idea involves limiting the area to be measured. This is a feasible option because having the 3D shape of a particular part of the cloth is often enough for many applications. The proposed method can be easily adapted to perform the 3D measurement for multiple local areas.

The next phase of this research entails implementing the cloth handling application in [4] by using the proposed method.

REFERENCES

Fig. 14. Predicted blocks before the improvements of Section 4.C

Fig. 15. Predicted blocks after the improvements of Section 4.C

Fig. 16. Predicted blocks and dissimilarities by using the area-average method