社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－CVIM－156（15）
2006／11／10

An Active Method of 3-D Distance Measurement Using An Analysis Map

Jonghoon WON[†]　　Hiroshi NAGAHASHI[‡]

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology[†]

Graduate School of Science and Engineering, Tokyo Institute of Technology[‡]

We propose an active way of 3-D distance measurement for scene perception. A real-world environment is composed of various kinds of objects. The most adequate parameters of a measurement system may be different according to the purposes of measurement and the characteristics of objects. It needs to adjust the system parameters such as camera position, pan and tilt angles, focal length, zoom and so on. Firstly, we perform 3-D distance measurement of the whole scene by parallel stereo vision. Next, we perform scene analysis by considering two kinds of data, i.e., depth and texture, to plan a purpose-oriented measurement strategy. Some regions are selected for targets and then their positions are estimated from the measurement results of the whole scene. After adjusting the camera parameters to a target region, a fine measurement is performed again.

1 Introduction

Scene perception supports many applications in computer vision, such as intelligent robots [1], auto-navigations [2], support systems for handicapped persons [3] and so on. The processes in scene perception can be divided into scene partitioning, analysis and recognition. The distance measurement of a real-world environment is one of the fundamental problems in each process and has been studied by many researchers. The applications mentioned above have a lot of concerns with human life. The distance measurements for scene perception may be performed efficiently by considering the human vision. We discuss a measurement method for scene perception in the following, while focusing on human behaviours.

Human eyes have a wide field of vision. However, Humans can perform a perception only in a small region of the whole scene at one time, when they observe a certain real-world environment. It can be said that it arises from two reasons. Firstly, there is the limitation of sensing ability in human eye. The images obtained through the eye become blurry by going toward the periphery of visual field. Only a small region is sensed fine. Secondly, it is due to the limitation of recognition ability. Generally, there are various kinds of objects in a real-world environment. They can't recognize all objects in a scene only by one gaze. Strictly speaking, they can recognize just one object at a time. Most measurement systems using stereo configurations similar to human eyes also have the same restriction with a human being. To compensate the two limitations mentioned above, humans scan a scene by moving their gaze from a place to others. Measurement systems can overcome these limitations by controlling camera parameters, such as view point, vision field, measurement position and so on. Which strategy is best for scene perception in the control of camera parameters? In the following, we discuss this issue, while considering human vision.

A real-world environment is composed of various kinds of objects. Humans observe surroundings, while moving their glances in order of the area with a high interest. A scene can be divided into various kinds of regions based on the degree of attention by the purpose of measurement. Some regions may be not necessary to measure and others may be necessary to measure accurately. We can perform the distance measurement efficiently by adjusting the measurement method according to the attention degree of each region. There are many researches that perform the distance measurement by controlling various camera parameters [4]. However, in most researches, the issues specific to the attention degree haven't been considered.

In this paper, we propose an active way of 3-D distance measurement for scene perception that can select adequate strategy in each region. To construct a measurement strategy, it is necessary to estimate the attention degree of each region. We perform

scene analysis to estimate the attention degree of each region. We will also address how to estimate it from the result of scene analysis.

The remainder of the paper is organized as follows. Section 2 describes how to perform a measurement for scene perception. Section 3 explains our active camera system. Section 4 describes how to estimate an attention degree and how to decide a target region. Section 5 describes how to control the various camera parameters and how to measure a target region. Section 6 presents some experiments.

2 Overview of the measurement

The stereo matching is one of the most common methods in measuring 3-D distance and is the most similar to human stereopsis. In stereo matching, finding corresponding points is the most difficult step. Many researches have focused on the correct correspondence. Scharstein and Szeliski [5] presented taxonomy of dense, two-frame stereo methods. Looking at the results of their experiments, global optimization methods based on 2-D MRFs [6][7] generally perform the best in all regions of the image (overall, textureless, and discontinuities) and local methods [8] perform less well. However, global methods are too slow for practical use. Felzenszwalb and Huttenlocher [9] presented new algorithmic techniques that substantially improve the running time. It is a belief propagation approach, which is categorized in the global optimization methods. We implement the correspondence matching using the method proposed by Felzenszwalb et al.

In the stereo matching, the camera parameters such as viewpoint, focal length and position, are usually fixed so that the available region of measurement is restricted. As mentioned in section 1, for scene perception, we perform the measurements while adjusting the camera parameters to some targets. The parameter planning for measurements has been discussed in a lot of researches [10][11]. However, there are few researches for scene perception. It is discussed below which planning is better for scene perception.

In our measurement system, we use a camera that can control zooming. Using the function of zooming, we can regulate the resolution of a target region. We perform the distance measurement, in an adequate resolution based on the attention degree.

For measuring a target after zooming in, it needs to maintain placing a target in a range of view. There are a number of camera positions and viewing directions in which the above constraint is sufficed. Which direction and position are best to measure a target?

Firstly, we will address a viewpoint problem. As regulating resolution higher by zooming in, the range of view becomes smaller. If a target is kept placed in the range of view and is positioned in the center of image plane, the zoom factor can be controlled to the maximum magnification. Next, we consider the problem of viewing direction. A projected image becomes different if the viewing direction of the camera changes. The measurement results of target regions are desired to be integrated into the whole scene. If there is no change in viewing direction, the results with various resolutions can be integrated easily.

Considering the constraints mentioned above, we have constructed the measurement planning. The processes of our measurement are mentioned briefly below. We suppose that measurement purpose is to adjust resolution to each region. Each process is described minutely in sections 4 and 5.

- Measure the whole scene on the parallel stereo vision and perform the scene analysis.
- Decide a target region and evaluate camera parameters adequate for the measurement of the target region.
- After adjusting the viewpoints of two cameras to the target region, verify the results and refine.
- Remeasure the target region after regulating resolution.

3 Active camera system

3.1 System overview

Our system based on the proposed method is composed of one turn-table, one xy-axis stage and two CCD-cameras (Fig .1). The rotating range of the turn-table is 360^{o} and the single-pulse rotation angle is 0.05^{o}. On the xy-axis stage, the range of travel is 200mm respectively and the single-pulse travel is 4μm. The both cameras can control the pan, tilt angles, and the focal length. The range of pan is $\pm 100^{o}$. The range of tilt is $\pm 19^{o}$. The focal length is from 5.4mm to 64.8mm. The turn-table is joined up to the base plate of the xy-axis stage, while centering the axis of rotation on the base plate of x-axis. The CCD cameras are integrated with 150mm baseline as shown in Fig.1 (a). The camera system is joined up to the turn-table while centering the rotation axis of turn-table on the middle of the line connecting between the focal points of the cameras. The stages are controlled by commands sent from a PC via a RS232C interface and also the two CCD cameras are controlled directly by PC.

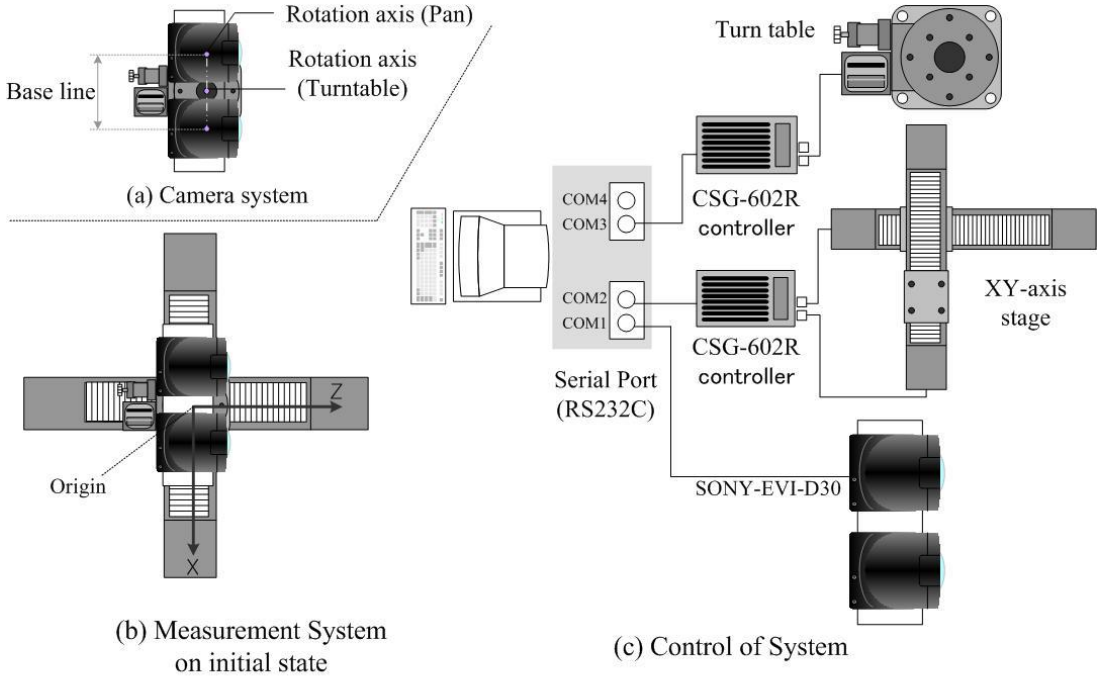


Fig.1 Active Camera SYSTEM

3.2 Coordinate systems

The initial state of the measurement system is shown in Fig.1 (b). In the initial state, the rotation axis of turn-table is centered on each axis in the xy-axis stage and there is no rotation in the camera system. In this paper, we use 3 kinds of coordinate systems. Firstly, we describe a world coordinate system ($X_wY_wZ_w$). Its origin O_w is the middle point of the line drawn between two focal points of the cameras in the initial state. Each axis is also described in the initial state. The X_w-axis is parallel to the line connecting between each focal point of the cameras. The Z_w-axis is parallel to the optical axis of the camera. The Y_w-axis is pointing out of the page. The world coordinate system is fixed independent of controlling various parameters.

Secondly, a camera coordinate system is explained. The camera coordinate system ($X_cY_cZ_c$) is the same with the world coordinate system in the initial state. However, the X_c-axis, Z_c-axis, and O_c are changed by controlling the turn-table or xy-stage as like shown in Fig.1 (b).

Thirdly, in a image coordinate system, the x-axis is parallel to the horizontal on the image and the y-axis is parallel to the perpendicular on the image. The origin is the center of image. The image coordinate system is based on pixel coordinates.

4 Target decision

4.1 Analysis map

For the scene analysis, we use two kinds of factors, i.e., disparities $d(i, j)$ and colors $c(i, j)$. $d(i, j)$ are estimated by performing the parallel stereo matching in a whole scene. As mentioned in section 2, we adopted the method proposed by Felzenszwalb et al, to solve the correspondence problem. We perform the scene analysis to estimate the attention degree of each region. In this paper, the attention degree is decided by estimating how much more information is acquired through regulating resolution. The scene analysis may be performed efficiently by using a difference image formed by subtracting the RGBs of original image from those of a slightly different resolution image. There are several methods to form an image with different resolution such as, controlling the zoom function, averaging neighbors around a pixel. Defocusing has a similar impact and is performed easily. We calculate $c(i, j)$ by $(|r'(i, j) - r(i, j)| + |g'(i, j) - g(i, j)| + |b'(i, j) - b(i, j)|)/3$. Here, $r(i, j)$, $g(i, j)$ and $b(i, j)$ are RGB values of pixels in the original image $p(i, j)$. $r'(i, j)$, $g'(i, j)$ and $b'(i, j)$ are RGB values of pixels in the defocused image $p'(i, j)$. At first, we calculate the variances of disparities $d_v(i, j)$ and RGBs $c_v(i, j)$ respectively, within the prepared

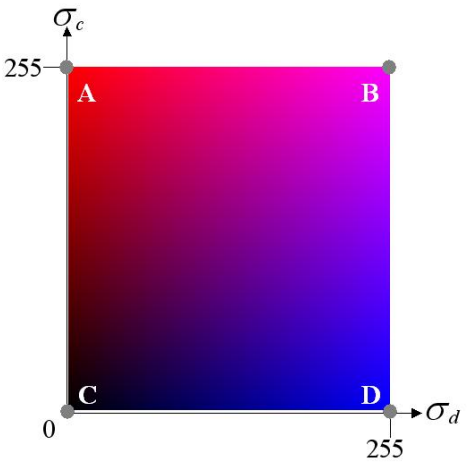


Fig.2 Labels used for scene analysis: The labels are expressed by colors of r and b whose values give σ_c and σ_d

window of size $[W_s=(2w_i+1)\times(2w_j+1)]$.

$$d_v(i,j)=\frac{1}{W_S}\sum_{l=-w_j}^{w_j}\sum_{k=-w_i}^{w_i}(d_a(i,j)-d(i+k,j+l))^2 \quad (1)$$

$$c_v(i,j)=\frac{1}{W_S}\sum_{l=-w_j}^{w_j}\sum_{k=-w_i}^{w_i}(c_a(i,j)-c(i+k,j+l))^2 \quad (2)$$

$d_a(i,j)$ and $c_a(i,j)$ are the averages of disparities and RGBs within the window respectively. Next, we calculate the standard deviations of $d_v(i,j)$ and $c_v(i,j)$ respectively. The both variances are trimmed by each standard deviation and linearly scaled in [0,255]. Fig.2 shows the labels diagram which is 2-D plane of (σ_d, σ_c). The labels (σ_d, σ_c) are expressed with red and blue components of RGB representation on σ_c-axis and σ_d-axis respectively. There are 256×256 kinds of labels. The high σ_d of the labels diagram means that depth changes rapidly in a real world. Similarly, the high σ_c means that texture changes rapidly.

4.2 Scene analysis and target decision

All pixels of whole scene image are labeled on the labels diagram by σ_c and σ_d. The labeled image is named as an analysis map. By using the analysis map, it is predicted that texture is changed rapidly in a region which is composed of pixels with $L(\sigma_d, \sigma_c)$ close to **A**. In such region, we can acquire more information about texture by regulating resolution high, however, it isn't necessary to perform distance measurement again. In the case of **B**, texture and disparity are changed rapidly. We can acquire more information of both texture and depth by regulating

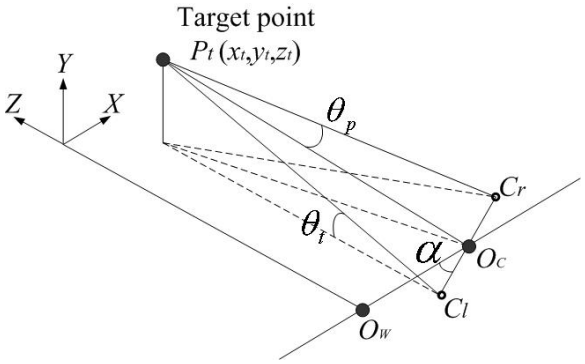


Fig.3 Control of parameters: (x_t, y_t, z_t) is the coordinate of the target point in the camera coordinate system

resolution higher. In the case of **C**, it is textureless. In such region, we can perform measurements by enlarging the size of searching window. In the case of **D**, the possibility that correspondence error is occurred becomes higher, because the change of depth is rapidly in spite of the fact that it is textureless. Such region must be remeasured.

Label that characterizes a high attention degree may be changed into either of **A**, **B**, **C** or **D** according to the measurement purpose. The target region can be decided by putting the constraints D_T and C_T on each axis σ_d and σ_c respectively.

5 Measurement of target regions

5.1 Control of camera parameters

To compute the camera parameters, a target point is selected in a target region [12]. The target point is decided by averaging all coordinates of pixels in the target region. We describe below how to adjust the camera parameters, considering the constraints mentioned in section 2.1. All processes are divided as follows.

1. Align the optical axis of left camera in front of the target point by translating the camera system in X-direction of the xy-axis stage. (In case target image is left one.)
2. Control the pan and tilt of both cameras and turn-table stage to focus on the target point.
3. Control the zoom factors of cameras.

In the process 1, the amount of shift is computed easily. The target point (x_t, y_t, z_t) have already been calculated from the measurement result of scene and the baseline b is known to be 150mm. The camera system is shifted by Eq. (3). In Eq. (3), ε is the gap between the x-coordinate of C_l and x_t caused by the rotation of turn-table. ε is computed from $(b\text{-}b\cos\theta_p)/2$. θ_p is the pan angle. In the process 2, The turn-table is rotated by α calculated in Eq.(4).

$$T_x = (b/2 + x_t) + \varepsilon \quad (3)$$

$$\alpha = \sin^{-1}\left(\frac{b}{(2z_t + b\sin\theta_p)}\right) \quad (4)$$

$$\theta_t = \tan^{-1}\left(\frac{2y_t}{(2z_t + b\sin\theta_p)}\right) \quad (5)$$

$$\theta_p = \tan^{-1}\left(\frac{b}{2z_t}\right) \quad (6)$$

In Fig.3, O_w, O_c are the origins of the world coordinate system and camera coordinate system respectively. C_l and C_r are the focal points of the left and right cameras. The tilt angle θ_t and the pan angle θ_p are calculated respectively in Eq. (5) and Eq. (6). To solve from Eqs. (4) to (6), it needs to calculate θ_p and α at first. The pan angle is equal to the rotation angle of the turn-table. θ_p and α are calculated by solving the simultaneous equations of Eq.(4) and Eq.(6). θ_p is just the amount of pan angles of both cameras, and has a positive sign in the left camera and a negative sign in the right camera. In the process 3, the magnification N of the zooming is obtained as follows.

$$N = \frac{S^2}{\lambda s^2},\ s = \max(w,h),\ S = \min(W,H) \quad (7)$$

Here, W and H denote the width and height of the image plane. w and h denote the width and the height of the target region. λ is a constant that is a little larger than 1. The target region will be projected within the image plane of both cameras after zooming in by using Eq.(7).

5.2 Verification and refinement

After adjusting the camera parameters, the target point will be positioned on the center of image, if the adjustment of parameters and the result of measurement are accurate. We investigate how much the target point is swerved from the center of image by comparing the correlation between the target region and candidate area in the image that is projected after adjusting parameter. We can calculate the swerved distance E_d on X-axis of world coordinate system from Eq.(8).

$$E = e_x z_t / f \quad (8)$$

Here, e_x is x-coordinate of target point in left mage coordinate system. z_t and f are Z-coordinate of the camera coordinate system, and focal length respectively. We can estimate not only E_d but also real 3-D coordinates (x_t, y_t, z_t) of target point by calculating the equation mentioned above recursively.

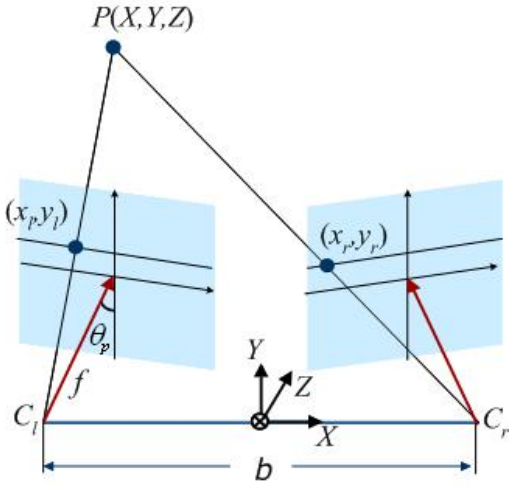


Fig.4 Non-parallel stereo

5.3 Extraction of 3D-Coordinates

In the parallel stereo matching, the correspondence search of a point is performed just on the row with the same height in reference image, as a target point and its reference point locate on the same scan line of the both images. In our method, the above constraint is not adopted, as optical axes of cameras are not always parallel. We have derived some equations for the correspondence search and the extraction of 3D-coordinates in non-parallel [12]. The correspondence search is performed on the epipolar lines that are derived with the constraint that the both cameras are controlled with the same amount of angle.

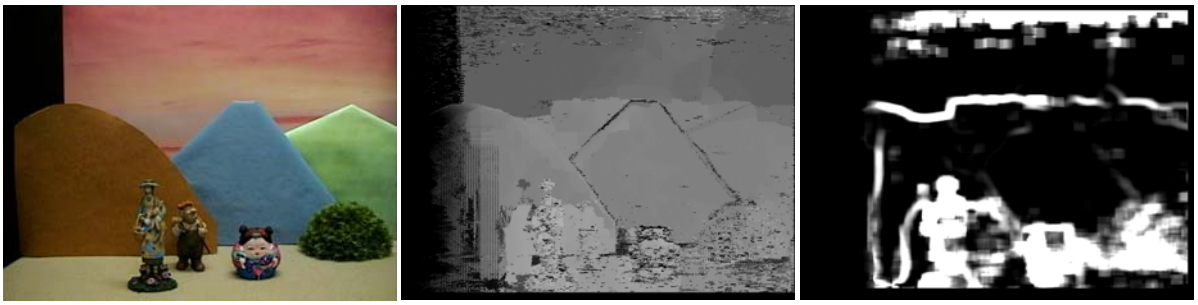
$$y_r = \left(\frac{x_r}{x_r'}\right)\left(\frac{x_l'}{x_l}\right) y_l \quad (9)$$

$$x_l' = \frac{x_l f}{f\cos\theta_p - x_l \sin\theta_p},\ x_r' = \frac{x_r f}{f\cos\theta_p + x_r \sin\theta_p}$$

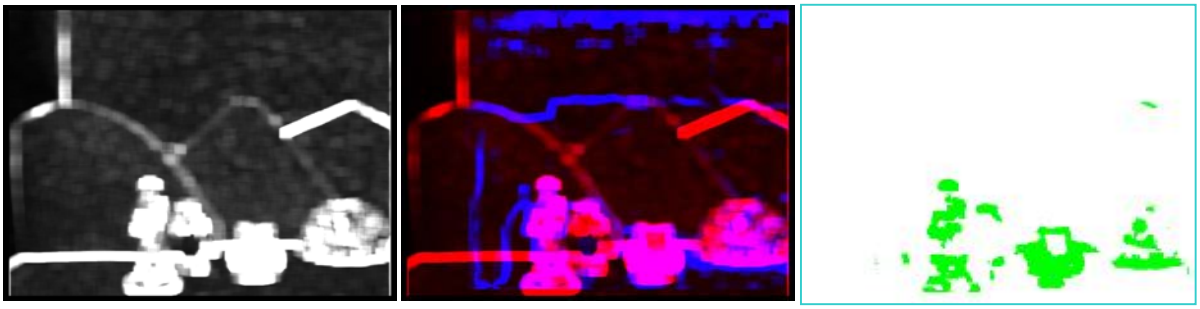
A point $P(X,Y,Z)$ in the 3D space will be projected respectively on the left and right 2-D image plane. (x_l, y_l) denotes the coordinate of the projected pixel in the left image coordinate system. (x_r, y_r) denotes one in the right image coordinate system. After searching the corresponding pixels of pixels in a target image, 3-D distances are computed from Eq. (10). The focal length f, the base line b, and the pan angle θ_p are known. The focal length is computed by the calibration of the camera.

$$X = \frac{b(x_l' + x_r')}{2d'},\quad Y = \frac{Z y_l}{f\cos\theta_p - x_l \sin\theta_p} \quad (10)$$

$$Z = \frac{bf\cos\theta_p}{d'},\quad d' = 2f\sin\theta_p + x_l' - x_r'$$



(a) Left image of whole scene (b) Disparity map of whole scene (c) Variances of $d(i,j)$



(d) Variances of $c(i,j)$ (e) Analysis map (f) Target regions

Fig. 5 Outdoor scene: toy world

6 Experimental results

We show some experimental results to verify the effectiveness of the method proposed in this paper. All of the image size is 640x480 in this experiment. We prepared a toy world similar to a real-world environment (outdoor). The size of the toy world is 850×850(mm). Our measurement system is positioned in front of the toy world. Fig. 5(a) shows the left image of the whole scene. We performed the measurement of the scene. The disparity map of the whole scene is shown in Fig. 5(b).

6.1 Scene analysis

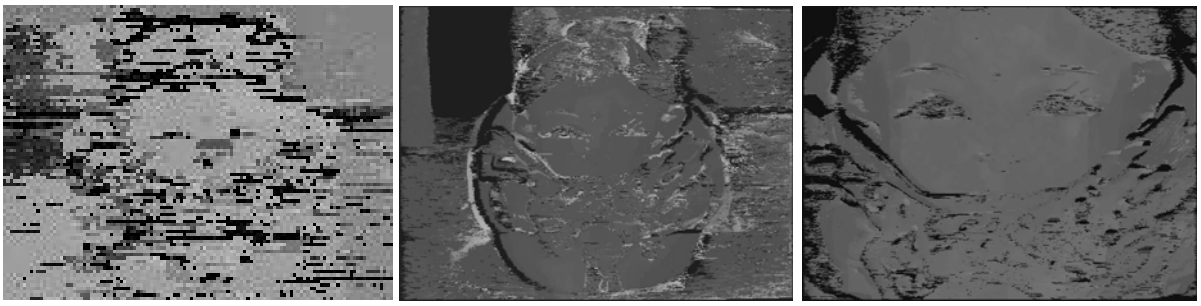
We performed the scene analysis of the scene to show the effectiveness of the proposed method. Fig. 5(c) and (d) show the images generated respectively by expressing each variance with intensity value. The analysis map is shown in Fig. 5(e). We can see easily that the four kinds of regions were expressed well, which is mentioned in section 4. The property near label **A** appears strongly in the edges of objects. The region where is textureless strongly shows property near label **C**. It is clearly shown that there are some regions with "blue" labels (featured by **D**) where mismatchings of correspondence have occurred. The regions of dolls show the property near label **B**, where a lot of variations of the colors and depths exist.

6.2 Decision of targets

The analysis map of the scene was shown in Fig. 5(e). The target regions were extracted by setting the both D_T and C_T in [200,255]. The extracted regions are shown in Fig. 5(f). There may be some regions which are composed of a few pixels in the extracted regions. There is high possibility that such regions were caused by miss-extraction. Only the regions with more than κ pixels are taken as target regions. The target point is decided by averaging all coordinates of pixels in the target region. Four target points were selected, given κ=200 and marked by ⊗.

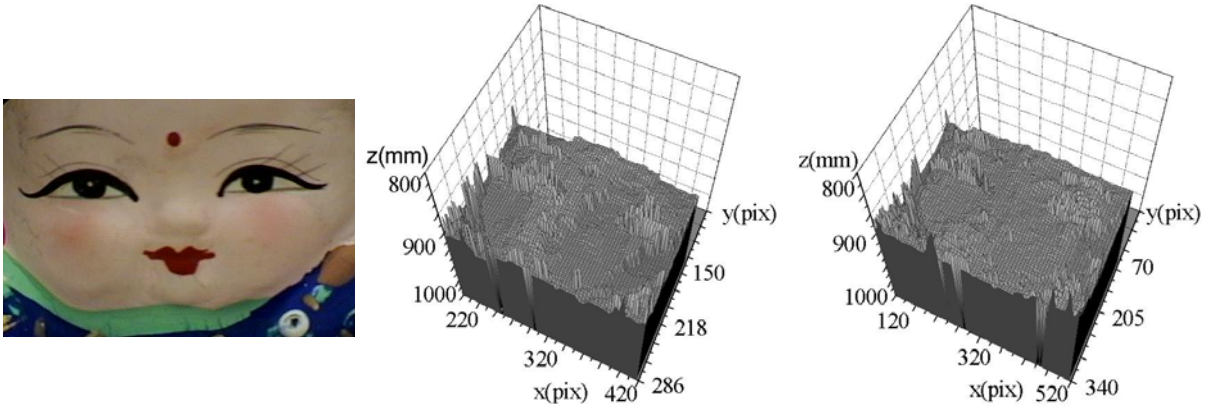
6.3 Control of parameters

The viewing directions of the both cameras are controlled to locate the target region on the center of each image. The amounts of each rotation and translation are calculated from Eqs. (4) to (6). In this time, the result of the target point P_t (417,404) is shown. The 3-D coordinate of the target point is calculated from Eq. (9). Firstly, the camera system was translated by Eq. (3). Next, it was refined by Eq. (8). After adjusting the camera position and viewing direction, the resolutions of both images are regulated high with the magnification calculated by Eq. (7) (N=6).



(a) Magnification, x1 (b) Magnification, x6 (c) Magnification, x12

Fig. 6 Disparity Maps of Target Region with Different Resolutions



(a) Region of comparison (b) Magnification, x6 (b) Magnification, x12

Fig. 7 Comparison of the results in 3-D

6.4 Distance measurements

Fig.6 (a) shows the disparity map of target region that is the part of the results of the whole scene. Fig. 6(b) shows the disparity map of target region after controlling resolution according to calculated magnification. Fig. 6(b) shows more detailed result than Fig. 6(a). In Fig. 6(b), the outline of doll is distinguished from the surroundings and occlusion is shown clearly. We can see also easily that the rate of mismatching is lower in Fig. 6(b) than in Fig. 6(a). Fig.6(c) shows the disparity map acquired by regulating only the resolution to the maximum. The shape of face was more accurately acquired in higher resolution. In Fig.7, the measurement results are expressed in 3-dimension to show the effectiveness of resolution control clearly. Fig. 7(a) shows the region that is compared. Fig. 7(b), (c) are respectively the partial results of Fig. 6(b), (c). In Fig. 7(b), (c), x and y-axis denote respectively x and y coordinates of image coordinate system and z-axis denotes Z coordinate of world coordinate system. We can see that the shape of face was expressed more smoothly in Fig. 7(c) than in Fig. 7(b) especially in chick and forehead.

7 Conclusions

Scene perception has a lot of concerns with most of researches in computer vision and the 3-D distance measurement is one of the fundamental problems in scene perception. Although many researchers have focused on the 3-D distance measurement, there are few measurement methods for scene perception. We presented an active measurement method of the 3-D distance, which is useful for scene perception. There are various kinds of objects in a real world environment. The most adequate strategy of measurement is different according to each region of the whole scene. We can perform the 3-D distance measurement efficiently by constructing the adequate strategy for each region respectively.

To estimate the attention degree, we performed the scene analysis by using the analysis map. The efficiency of the analysis map was shown well in the experimental results. The target region was extracted by the results of scene analysis and remeasured fine after regulating resolution high. In this paper, we aimed to perform the remeasurement of the regions in which textures and disparities are changed rapidly. In other words, the target region may be composed of

pixels with the labels of (σ_d, σ_c) close to B. Other 3 kinds of regions characterized with **A**, **C**, or **D** can be also adopted for target region by regulating the D_T and C_T. We divided whole scene into four kinds of regions and proposed the measurement strategies of each region briefly. We will also try efficient measurements of other 3 kinds of regions, and devise the more detailed method of scene analysis. An integration of measurement results from different camera parameters is also our future works.

[References]

[1] Breazeal, Cynthia, A. Edsinger, P. Fitzpatrick and B. Scassellati : ``Active Vision for Sociable Robots", IEEE Transactions on Man, Cybernetics, and Systems, Part A: Systems and Humans, Vol. 31,No. 5,pp. 443-453 (2001).

[2] Luke Fletcher, Nichoals Apostoloff, Jason Chen, Alexander Zelinsky : ``Computer Vision for Vehicle Monitoring and Control", Proc. 2001 Australian Conference on Robotics and Automation (2001).

[3] Yoshihiro Kawai, Makoto Kobayashi, Hiroki Minagawa, Masahiro Miyakawa, and Fumiki Tomita : ``A Support for Visually Impaired Persons Using Three-Dimensional Virtual Sound", ICCHP 2000, pp.327-334 (2000).

[4] W. Fellenz, K.Schluns, A. Koschan, and M. Teschner :``An Active Vision System for Obtaining High Resolution Depth Information", CAIP'97, pp.726-733 (1997).

[5] Daniel Scharstein, Richard Szeliski : ``A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspo- ndence Algorithms", In IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 131-140 (2001).

[6] Jian Sun, Heung-Yeung Shum, Nan-Ning Zheng: CIE Div. 8 TC -07 : ``Stereo Matching using Belief Propagation", ECCV 2002, pp. 510-524 (2002).

[7] J.L. Barron, D.J. Fleet, S.S. Beauchemin : ``Performance of Optical Flow Techniques", International Journal of Computer Vision, Vol. 12. No. 1, pp. 43-77 (1994).

[8] R. Zabih, J. Woodfill : ``Non-parametric Local Trans- forms for Computing Visual Correspondence", Third European Conference on Computer Visionpp. (1994).

[9] Pedro F. Felzenszwalb and Daniel P. Huttenloocher : ``Efficient Belief Propagation for Early Vision", IEEE Conference on Computer Vision and Pattern Reco- gnition. (2004).

[10] W. Brent Seales : ``Measuring Time-To Contact Using Active Camera Control", CAIP1995, pp. 944-949 (1995).

[11] Peter Lehel, Elsayed E. Hemayed, Aly A. Farag : ``Sensor Planning for a Trinocular Active Vision System", CVPR'99, pp. 306-312 (1999).

[12] Jonghoon WON, Ken'ichi MOROOKA and Hiroshi NAGAHASHI : ``Distance Measurement Of A Real-World Environment Using An Active Camera System", International Workshop on Advanced Image Technology, pp. 163-168 (2006)