8—1

Development of the Vision System for the HRP-2P Humanoid Robot

Yoshihiro Kawai*, Yutaro Fukase**, Ryohei Ikeno***, Yutaka Ishiyama***, Fumiaki Tomita*
*National Institute of Advanced Industrial Science and Technology (AIST)
**Shimizu Co. ***Stanley Electric Co.

Abstract

The present paper describes the 3D vision system of the HRP-2P humanoid robot developed by the Humanoid Robotics Project of the Ministry of Economy, Trade and Industry (METI) of Japan. HRP-2P was developed to work in conjunction with humans on uneven surfaces and to act autonomously. A number of high-level 3D computer vision technologies were required in order to accomplish these tasks. The present paper introduces the hardware of the vision system and a number of vision functions that were added to the Versatile Volumetric Vision: *VVV* system, a total 3D vision software package under development at AIST.

1 Introduction

The present paper describes the 3D vision system of the HRP-2P humanoid robot developed by the Humanoid Robotics Project of the Ministry of Economy, Trade and Industry (METI) of Japan[1][9]. Honda and Sony have both developed humanoid robots, ASIMO[10] and SDR-4X[11], respectively. However, these robots are primarily designed for entertainment purposes and are not practical as working robots. HRP-2P was developed to work together with humans on uneven surfaces and to act autonomously. Accomplishing these tasks requires accurate information of the 3D environment. However, a number of constraints exist for the vision system with regard to the design of the face and the size and weight of parts used in the head. This paper introduces the hardware of the vision system and a number of vision functions added to the Versatile Volumetric Vision: *VVV* system[8], and reports the demonstration at the ROBODEX2002 robot exhibition.

2 Hardware for HRP-2P

2.1 HRP-2P

Figure 1 shows an external view of HRP-2P. HRP-2P has a height of 154 *cm* and a weight of 58 *kg*. HRP-2P has 30 degrees of freedom, including two waist axes, so that HRP-2P is able to walk through narrow paths and on uneven surfaces. Moreover, HRP-2P can stand up after having fallen down. An improved HRP-2 will be released at the end of 2002.

2.2 Stereo camera system for HRP-2P

Figure 2 shows (a) head parts covered by a shield and (b) a stereo camera system. This system is usually covered by a shield, integrating the camera system into the HRP-2P design. The neck of HRP-2P is constructed to have pan and tilt functions, and the vision sensor consists of a stereo camera system to which three cameras are mounted. Two horizontal cameras are separated by 120 *mm*, and the third camera is located 60 *mm* above the two horizontal cameras. Three cameras are used in order to increase 3D measurement precision and reduce stereo correspondence errors. The structure of the system is simple, as a result of limitations that include a weight limit of less than 1 *kg*, and it have to be stable for vibrations. A first prototype which has a convergence function for each camera was developed. However, in the prototype, captured images were blurred by vibrations due to the light weight required for the rotation mechanisms. A camera lens having controllable zoom, focus, and iris functions were not used because of the weight limit. A wide-angle lenses having a single focal length (8 *mm*) was used to capture distant and nearby scenes simultaneously. Shutter speed is variable and is adjusted automatically under any lighting situation. The total weight is less than 700 *g*.

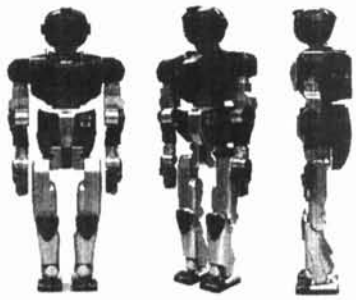
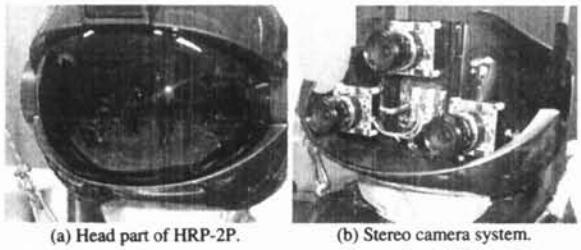


Figure 1: HRP-2P humanoid robot.



(a) Head part of HRP-2P. (b) Stereo camera system.

Figure 2: Stereo camera system for HRP-2P.

3 Vision functions

Our AIST group is developing an advanced 3D vision system called *VVV* system[8], which can be used for various purpose in many fields. The basic processes of 3D vision are camera calibration, range sensing of a 3D scene[2], describing the 3D shape of objects in the scene, and recognizing the 3D position and orientation of objects by matching with object models[6]. In addition, moving objects can be tracked by repeating these processes[7]. The relationship between HRP-2P and

*Address: AIST Tsukuba Central 2, Tsukuba 305-8568 Japan. E-mail: y.kawai@aist.go.jp

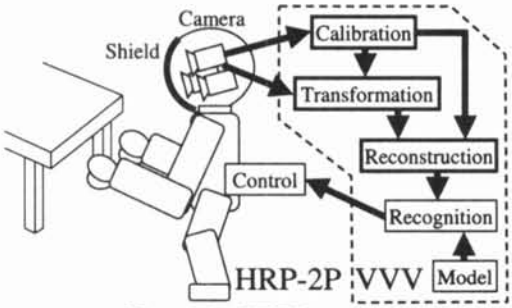


Figure 3: VVV system.

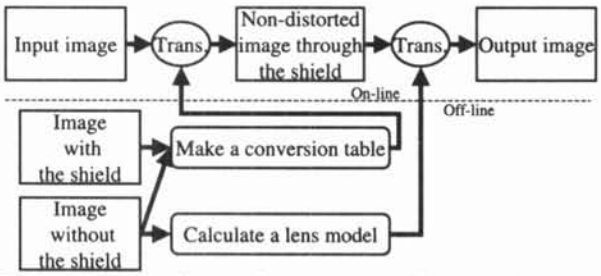


Figure 4: Flowchart of correction of distorted image.

the VVV system is shown in Figure 3.

This section describes four functions that were added to the *VVV* system: distortion correction for both a shield and a wide-angle lens, a method of best baseline selection, and depth correction by planarity constraints. These functions are necessary in order to obtain more accurate 3D data so as to perform operations stably and autonomously. Finally, the model-based recognition process is described briefly.

3.1 Distortion correction

Two distortion corrections are needed due to the use of the shield and wide-angle lens. One lattice pattern plane, as shown in Figure 5(a), is used for both correction methods. Figure 4 shows a flowchart of the distortion correction. At the calibration stage (off-line), two images, with and without the shield, respectively, are captured for the pattern plane. A conversion table between the two images is calculated by comparing two images. The lens model is calculated without the shield. In on-line mode, first, the input image is translated using the conversion table in order to remove the influence of the shield. Then, the image is translated to a non-distorted image based on the lens model.

3.1.1 Distortion correction for a shield

The vision system is covered by a shield for aesthetic purposes, as shown in Figure 2(a). As a result, the captured images are distorted by the shield. Practically speaking, modeling the shield shape and the positional relationship between the shield and the camera is difficult. A realistic solution to this problem is to create a conversion table. The correction algorithm is as follows. First, the position of the center of gravity of each circle is calculated for each image. Using a large circle in the image center, point correspondences between two images with and without the shield are calculated. If the distortion is not large and the lattice distance is narrow, affine transformation is concluded within four neighborhood points between two images. Transformation for all points is performed by interpolation based on the affine transformation.

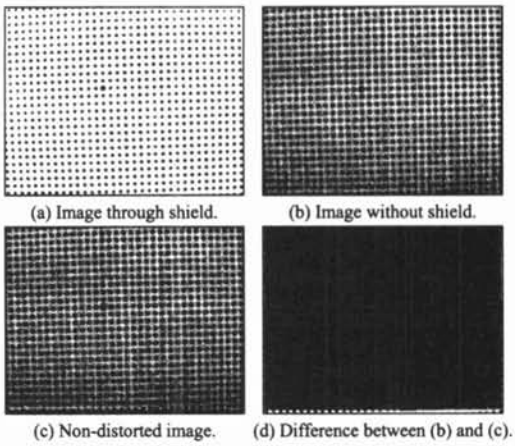


Figure 5: Results of distortion correction (shield).

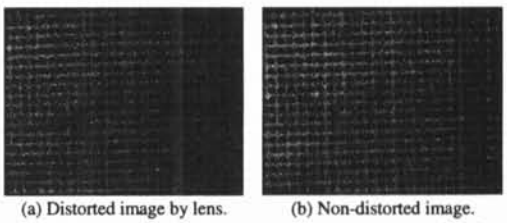


Figure 6: Results of distortion correction (wide-angle lens).

Figure 5 shows an experimental result obtained using this method. Figure 5(a) and (b) are captured images with and without the shield, respectively. The result of transformation to non-distorted image is shown in Figure 5(c), and the difference between Figures 5(b) and (c) is shown in Figure 5(d). This method is effective for shields with any shape.

3.1.2 Distortion correction for wide angle lens

Distortion of the lenses is then corrected using the image without the shield. Among the many distortion factors, the influence of the wide-angle lens distortion is the largest. Thus, at this stage, only the wide-angle lens distortion is corrected.

The lattice position P on a plane is expressed as follows:

$$P = i\vec{u} + j\vec{v} + \vec{w},$$

where i, j are integers, $\vec{u}, \vec{v}$ are non-parallel vectors, and $\vec{w}$ is the origin. The position of the large circle is defined as $\vec{w}$. Vectors for column and row directions are $\vec{u}, \vec{v}$. Any point can be expressed by the parameters (i, j). Data near the image center is generally only slightly distorted for the wide-angle lens, so the ideal positions without distortion for all lattice points are calculated using these data. The following equation is applicable between a real position (X', Y') and an ideal position (X, Y).

$$X' = X + (X - X_0)(k_1 r^2 + k_2 r^4),$$
$$Y' = Y + (Y - Y_0)(k_1 r^2 + k_2 r^4),$$

where $r^2 = (X - X_0)^2 + (Y - Y_0)^2$, and (X_0, Y_0) : is the cross point between the optical axis and the image plane. Four parameters k_1, k_2, X_0, and Y_0 are calculated by the correspondence between the real and ideal points. Usually one coefficient k_1 is sufficient to correct the distortion; however, more coefficients are needed for lenses of shorter focal length.

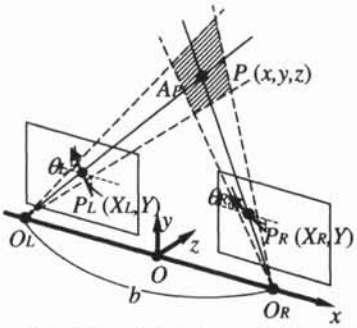


Figure 7: Evaluation of the best baseline under multiple baseline stereo.

Figure 6 shows an experimental lens distortion correction. Figure 6(a) shows a base image with lens distortion, an Figure 6(b) is this result converted to a non-distorted image using the lens model. The distortion is corrected, especially at the rim periphery. The error was less than 0.2 *pixels*.

3.2 Best baseline selection

In a multi-baseline stereo camera system, more accurate 3D information of horizontal segments can be reconstructed by combining inputs from other cameras. Although there exist several methods by which to acquire 3D data for a multi-baseline stereo system[4], we defined an evaluation equation based on the length of baseline, distance to target, and angle of epipolar line, as illustrated in Figure 7.

First, error values in the X direction at projected points P_L and P_R were calculated based on the angle θ, namely, $cos(\theta)$, because correspondence is more obscure at horizontal positions. Oblique area A_p, which is framed by the dotted lines, is the range of positions for corresponding points in the 3D space. A_p is proportionate to equation[5]

$$A_p \propto \frac{g(cos(\theta_L), cos(\theta_R))z^2}{bf},$$

where fuction $g()$ is the ambiguity function in the X direction, b is the length of baseline, and f is the focal length. In the HRP-2P vision system, f is very short, and $b \ll z$. Therefore, if f is shorter, then A_p becomes lager. A method by which to integrate the results on each baseline does exist; however, errors for non-best results may be larger in this situation. As a result, in this system, if there exists more than two correspondences while calculating 3D positions, only one result based on the baseline with the smallest A_p is selected.

Figure 8 shows an example of the best baseline selection. Figure 8(a) is a target scene, where a table is located approximately 80 *cm* distant from the cameras. Figure 8(b) shows the relationship between the three cameras as well as the epipolar lines and angles for the cameras. Reconstructed 3D data in each baseline is displayed in Figure 8(c). The selected lines are black, and the others are shown in gray. The 3D data reconstructed by baseline between Camera 1 and Camera 2 at a far edge is very noisy, because b_{1-2} is shorter than b_{0-1} and θ is almost 0. Little distortion is observed when using a pair of best lines.

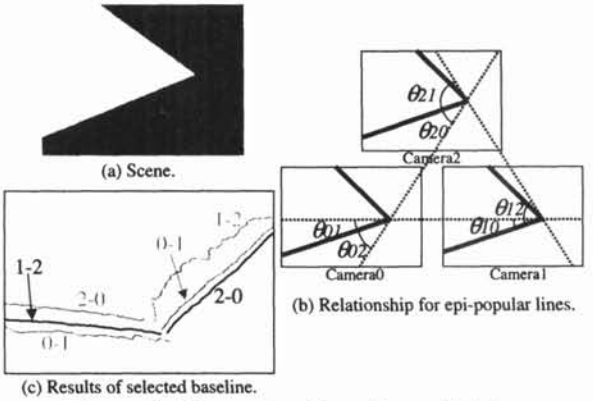


Figure 8: Example of baseline selection.

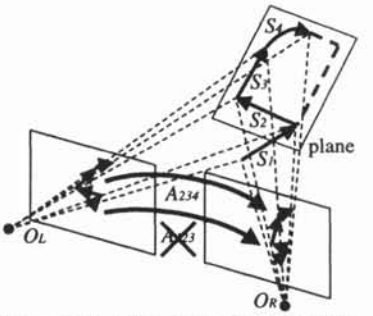


Figure 9: Planarity constraints.

3.3 Depth correction by planarity constraints

Three-dimensional data reconstruction is based on the segment-based method. The length of baseline is shorter than the distance to target due to the size and shape of the face of HRP-2P. As such, 3D accuracy, especially for depth, depends on the detection errors of the corresponding points. Consequently, simple 3D reconstruction using only depth information between corresponding points produces inadequate results. The depth can be refined and more accurate 3D structures can be reconstructed while maintaining the continuity of the 3D information by planarity constraints using the connectivity of the corresponding segments between images[3]. This reconstruction is based on the following theorem:

Theorem In a standard camera model, affine transformation between two projective images of any figures on a plane.

It is expressed as follows:

$$P_{R_jk} = AP_{L_ik} \quad (k = 1...n),$$
$$P_{L_ik} = \begin{pmatrix} X_{L_ik} \\ Y_k \\ 1 \end{pmatrix}, P_{R_jk} = \begin{pmatrix} X_{R_jk} \\ Y_k \\ 1 \end{pmatrix}, A = \begin{pmatrix} a & b & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where P_{L_ik} and P_{R_jk} are corresponding points. Confirming which points belong to a plane is easy. Figure 9 shows an example of planarity constraints. For segments (S_1, S_2, S_3) affine transformation A_{123} is not established, whereas A_{234} for (S_2, S_3, S_4) is established. The position of P_{R_jk} is transformed to P'_{R_jk} using A.

$$X'_{R_jk} = aX_{L_ik} + bY_k + c$$

This method is superior to the simple line fitting method, and solves the contortion problems and is

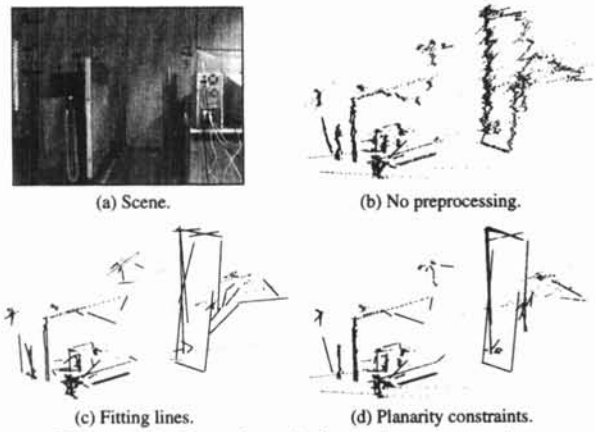


Figure 10: Results of planarity constraints.

adapted to not only straight lines, but also to curved lines on a plane.

Figure 10 is an experimental result of planarity constraints. In Figure 10, (a) shows a scene, in which a large panel is located at an approximate distance of 5 m, and (b) shows the result of 3D reconstruction by point correspondences only, in which the depth error is very large. In addition, Figure 10(c) shows the result of fitting straight lines for each straight line in each image. However, contortion problems occur on some vertices. Figure 10(d) shows the result using planarity constraints, which reveals less contortion. Refining 3D data by planarity constraints, 3D data at great distances is reconstructed comparative stably.

3.4 Object recognition

The 3D data are matched using object models in a database in order to identify what objects are present and determine their status.

In 3D Euclid space, a 4×4 transformation matrix T is defined by a 3×3 rotation matrix and a translation vector. In other words, recognition is the calculation of T by comparing the models and the data. This process consists of two phases, an initial matching phase and a fine adjustment phase. In the initial matching phase, T is roughly calculated by comparing the geometric features of the model and the data. Then, accuracy is iteratively improved using the entire set of data for candidates that are higher than the threshold. The best result is selected so as to correspond with the model. The details of this algorithm have been reported in a previous study[6]. An example of table recognition is shown in the next section.

4 ROBODEX2002

At the ROBODEX2002 robot exhibition[12] held in Japan, at which a number of companies and research institutions demonstrated their products, HRP-2P demonstrated its ability to recognize a table and carry it cooperatively together with a human operator. Figure 11(a) shows the detection of the table, and Figure 11(b) shows images captured at the edge of the table top. The results of recognition, with the model superimposed on 2D and 3D data, are shown in Figure 11(c). The grasping point was calculated to within an error of 2 mm based on the results of recognizing the table top edge. If the error was larger, carrying the table together with the operator would be unstable, because HRP-2P was under impedance control[9].

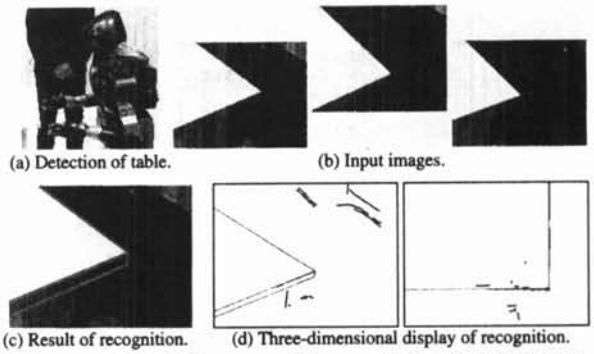


Figure 11: Table recognition at ROBODEX2002.

That is, HRP-2P was moved by pushing and pulling forces on the table generated by the operator. Correctly grasping the table is very important in order to accomplish this task. We were able to demonstrate the effectiveness of the stereo vision system in an exhibition for the general public.

5 Conclusions

The present paper describes the 3D vision system installed in the HRP-2P humanoid robot that enables the robot to work cooperatively with humans. A lightweight three-camera stereo system having auto iris control for use under any illumination was developed. In order to enable the robot to acquire high-accuracy 3D data, additional vision functions, including distortion correction, best baseline selection, and depth correction by planarity constraints, were developed and added to the *VVV* system. The effectiveness of the system was confirmed by the success of the HRP-2P demonstration at ROBODEX2002. In the future, we intend to expand the vision functions of the HRP-2P, adding functions such as walking path detection, in order to establish HRP-2(P) as a practical humanoid robot.

References

[1] H. Inoue, S. Tachi, Y. Nakamura, K. Hirai, N. Ohyu, S. Hirai, K. Tanie, K. Yokoi, H. Hirukawa, "Overview of Humanoid Robotics Project of METI", Proc. of Int. Symp. Robotics, pp.1478–1482, 2001.

[2] Y. Kawai, T. Ueshiba, Y. Ishiyama, Y. Sumi, F. Tomita, "Stereo Correspondence Using Segment Connectivity", Proc. ICPR'98, 1, pp.648–651, 1998.

[3] Y. Kawai, F. Tomita, "More Accurate 3D Reconstruction by Stereo Vision", Proc. of MIRU2002, 1, pp.159–164, 2002.[in Japanese]

[4] M. Okutomi, T. Kanade, "A Multiple-Baseline Stereo", IEEE Trans. PAMI, 15, 4, pp.353–363, 1993.

[5] J. J. Rodríguez, J. K. Aggarwal, "Stochastic Analysis of Stereo Quantization Error", IEEE Trans. PAMI, 12, 5, pp.467-470, 1990.

[6] Y. Sumi, Y. Kawai, T. Yoshimi, F. Tomita, "3D Object Recognition in Cluttered Environments by Segment-Based Stereo Vision", International Journal of Computer Vision, 46, 1, pp.5–23, 2002.

[7] Y. Sumi, F. Tomita, "Hyper Frame Vision: A Real-Time Vision System for 6-DOF Object Localization", Proc. of ICPR2002, III, pp.577–580, 2002.

[8] F. Tomita, T. Yoshimi, T. Ueshiba, Y. Kawai, Y. Sumi, "R&D of Versatile 3D Vision System VVV", Proc. of SMC'98, TP17-2, pp.4510–4516, 1998.

[9] K. Yokoyama, J. Maeda, T. Isozumi, K. Kaneko, "Application of Humanoid Robots for Cooperative Tasks in the Outdoors", Proc. of IROS2001, 2001.

[10] `http://world.honda.com/robot/`

[11] `http://www.sony.net/SonyInfo/News/Press/2j00203/02-0319E/`

[12] `http://www.robodex.org/e/`