3D Segmentation using Interval Analysis and Pre-attentive Behaviour for a Humanoid Robot

Olivier Stasse*, Benoît Telle+, and Kazuhito Yokoi*

* *AIST/ISRI-CNRS/STIC Joint Japanese-French Robotics Laboratory (JRL),* + *3D Vision Group*
Intelligent Systems Research Institute (ISRI),
National Institute of Advanced Industrial Science and Technology (AIST)
AIST Central 2, Umezono 1-1-1, Tsukuba, Ibaraki, 305-8568 Japan
{olivier.stasse,kazuhito.yokoi}@aist.go.jp

***Abstract*— This paper presents a 3D object segmentation algorithm based on dense 3D map provided by a stereoscopic vision system. The novelty of this paper is to use Interval Analysis for deciding to which region a 3D point should be merged with. This algorithm is used to implement an exploration behaviour on the HRP-2 humanoid robot.**

***Index Terms*— 3D reconstruction, Interval Analysis, humanoid**

I. Introduction

This paper presents a straightforward application of Interval Analysis applied to computer vision. The main idea presented in [1] is to reformulate the projective camera model by modeling the pixel noise as an interval. Using this reformulation to solve the 3D reconstruction problem [2], the result is a bounding box in which lies with certainty the reconstructed point. The main theoretical developments have already been presented in [1] and then will only be recalled briefly in section V.

The application of this result presented in this paper is to aggregate 3D points of dense range maps in the Euclidean space. This provides potential targets for exploration to a mobile robot, or obstacles to avoid. Here we will stress the application for object exploration and propose an implementation on a humanoid robot.

The remainder of this paper is as follows: in section II the motivations of this work is presented, in section III the algorithm to compute the dense map is quickly presented. Section IV presents some remainder on 3D reconstruction. Section V briefly introduces the 3D reconstruction reformulation in Interval Analysis. Section VI explains the algorithm used for 3D growing region. Section VII presents the experiments realized with the HRP-2 humanoid robot.

II. Motivations

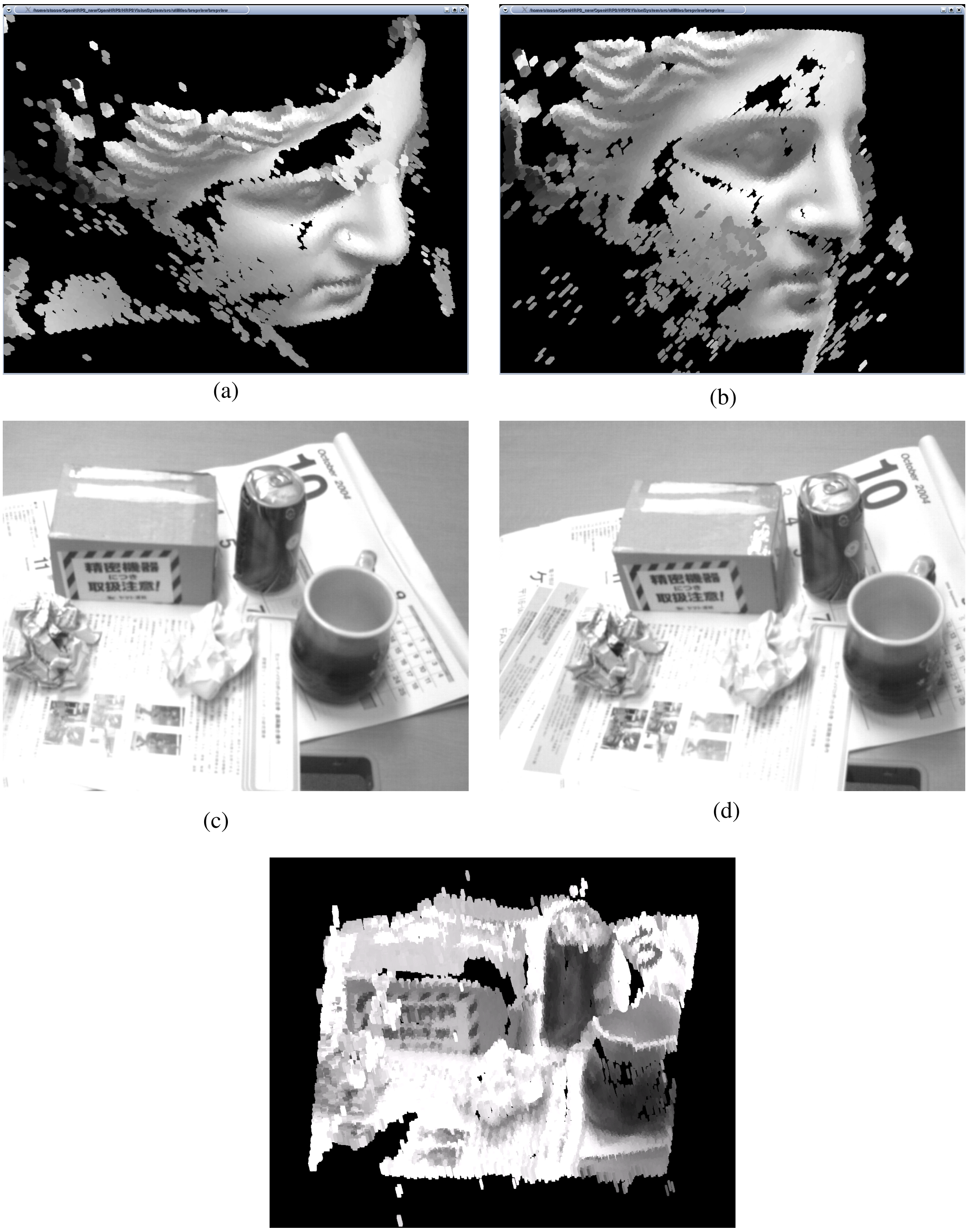
In order to increase the autonomy of a robot, it is necessary to develop some behaviour where the robot is able to detect an unknown object from the environment, and move towards to examine it. This is especially useful when the robot as to deal with objects for which he has not been programmed for, or in case of exploration. The main difficulty related to such behaviour is the assumptions needed to extract an *object* in the broad sense from the visual stream. The assumptions are usually build upon *appearance* or/and *geometry*.

A. Biologically inspired approach

Past works has been inspired by studies on biological vision systems, more precisely using the *visual attention* paradigm introduced originally by Treisman [3]. Also several works on robotics on this particular concept already exists [4][5][6], in this formalism we are more interested in the sensitivity to salient point in the environment called the *bottom-up* process, see [7] for a recent review. Using a psychological model, Driscoll and al. [4] described the implementation of a system which can pop out salient objects. In this model, the saliency of a pixel is determined by its difference to its immediate neighbours on each feature extracted from the image. The point the most salient on a local area across all the features is elected, and the process is reiterate on a wider region of the image until one point is finally elected. In this case the geometry taken into account is related to the image topology, and the appearance is defined by the choice of the features. We proposed a real-time parallel implementation of this algorithm for a humanoid robot [5]. In this case the geometry is also driven by the image but a log-polar sampling is used to decrease the complexity, while the appearance is provided by optical flow, and Gaussian filters up to the second derivatives. In this particular case, the log-polar sub-sampling makes difficult to have precise localisation of the object in the 3D space, and makes the subsequent algorithms too complex to handle. More recently, Minato and Asada [6] proposed to use a probabilistic approach for learning the appropriate parameters for a given set of filters. The filters themselves namely a 3×3 spatial filter and a colour filter are build upon a training period. In this context we would like to have the robot building itself the object's model.

B. Computer Vision approach

In computer vision, the salient points of an unknown object's are provided by the Harris detector [8]. It is widely used because of its robustness in detecting even across several camera viewpoints. Thus it is very often use to reconstruct 3D scene or objects [8] which then give the



(a) (b) (c) (d) (e)

Fig. 1

Depth Map of the Milo Venus (a-b) using Interval Analysis. The uncertainty is represented using boxes. A scene view reconstructed with two images.

geometric description of the object. However because it will provides several candidates around a region with a corner, an election mechanism is needed. For all this reason, the information provided by this descriptor is generally sparse, and needs several images before providing reliable information. This means other views, and for a robotic application, this will involve motion. Here, this is the visual information which should provide the first information for generating a possible motion. The same argument applies to recent descriptors proposed by Lazebnik [9]. Those descriptors are very good candidates for registration of several dense range maps and create a 3D representation of the object such as proposed in [10], but *after* the pre-attentive stage.

C. Pre-attentive behaviour

Considering a humanoid robot evolving in a 3D environment such HRP-2 [11], it is mandatory to have a precise 3D location of the candidate. For this reason in this paper, we present a pre-attentive behaviour based on 3D region growing apply to dense 3D map. The naive implementation of 3D region growing requires usually a distance and a threshold to decide or not if the point will be merged to a region. The usual drawback is the difficulty to find a threshold adequate to the object, to the environment, and to the condition of illumination. Lin and al. in [12] proposes to use intensity pixel as a distance and anisotropic and adaptive filtering to automatically find the threshold. The anisotropic filtering is modified to ensure convergence, and adapted to local property of the image. In this paper, as the dense map is given in the 3D space reconstructed from stereoscopic view, we use the Euclidean distance. The threshold problem is tackled by using the concept of uncertain point introduced by Telle and al. in [1]. In this work a 3D point is given by its center and a bounding box in which the point is certain to lie. Thus the aggregation is done simply by checking if two bounding boxes intersect. If this is the case the two points are merged. The main interest of this approach is the origin of the fusion. It takes roots into the geometry of the stereoscopic system, and for the 3D segmentation does not requires any manual adjustment.

III. Dense range map

In the remainder of this paper, the cameras are consider through their *projective* model obtained after a calibration process. As proposed in [2], a camera's projective matrix named $\mathbf{P}$ is defined by:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}| - \mathbf{RC}] \tag{1}$$

with $\mathbf{K}$ the 3×3 *intrinsic* parameters matrix, $\mathbf{R}$ the 3×3 *orientation* matrix of the camera, $\mathbf{C}$ the 3×1 *centre* position of the camera.

The dense map is constructed through the following pipe-line:

lens distortion rectification
→ image coordinates rectification
→ iso-luminance filtering
→ stereo matching

The lens distortion rectification is performed by a second order polynomial following the method described in [13]. The image coordinates rectification is done using the projective matrix. This allows comparing two pixels in the same coordinates system. Indeed pixels along the same epipolar line have the same value along the y-axis. The iso-luminance filtering is performed by sub-sampling the range of intensity value, and testing the immediate neighbourhood of a pixel. The stereo matching is performed by computing the absolute difference between two areas in the left and right images along the epipolar line. The best match between points in the left image and points in the right image is the one with the smallest difference. The result is used as the entry to 3D reconstruction process.

IV. 3D reconstruction

In order to help understanding the following section, we recall briefly how a 3D point can be reconstructed up to an arbitrary scale, once a matching between two points is realized. Considering a 3D point $\mathbf{Q}$ noted $\mathbf{Q}_h$in homogeneous coordinates, and $\mathbf{Q}_{nh}$ in non-homogeneous coordinates. Its projection $\mathbf{q}_l$ and $\mathbf{q}_r$ on respectively the left and right image are given by [2]:

$$\mathbf{q}_l = \mathbf{P}_l\mathbf{Q}, \mathbf{q}_r = \mathbf{P}_r\mathbf{Q} \tag{2}$$

Those two equations gives the following over-determined linear system:

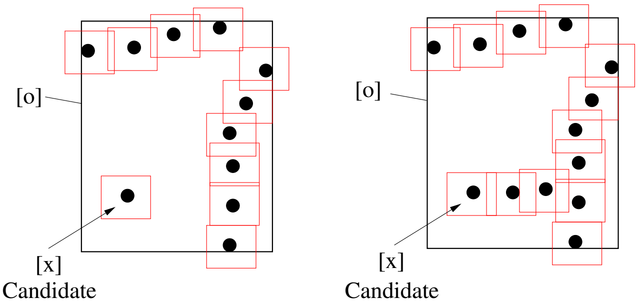
$$\mathbf{A}\mathbf{Q}_h = 0, \tag{3}$$

with

$$\mathbf{A} = \begin{bmatrix} q_l^0\mathbf{p}_l^{3T} - \mathbf{p}_l^{1T} \\ q_l^1\mathbf{p}_l^{3T} - \mathbf{p}_l^{2T} \\ q_r^0\mathbf{p}_r^{3T} - \mathbf{p}_r^{1T} \\ q_r^1\mathbf{p}_r^{3T} - \mathbf{p}_r^{2T} \end{bmatrix} \tag{4}$$

for which we have a total of four equations in four homogeneous equation. As the solution of this system is up to a scale, it is an over-determined system. Classically this system is solved by setting $\mathbf{Q}_h = (X, Y, Z, 1)$, and using the least-square method to solve this inhomogeneous equations.

In the following section we reintroduce briefly a new formulation of this problem allowing to find the 3D space in which the point $\mathbf{Q}$ lies with certainty.



(a)

(b)

Fig. 2

Candidates for merging : The point [**x**] is in the bounding box of the region, thus is a candidate. In (a), it can not be merged because it does not intersect any other points. In (b), it can be merged because it does intersect other points.

V. 3D reconstruction using Interval Analysis

In [1], a new camera model is introduced taken into account a different geometry of the pixel and its possible related error. This camera model is given by:

$$[\mathbf{q}] = E\left(\frac{\mathbf{P}\mathbf{Q}_h}{\mathbf{P}_3^t\mathbf{Q}_h}\right) + [\varepsilon] \tag{5}$$

Where E is the round operator which furnishes the nearest integer of a value. The denominator $\mathbf{P}_3^t\mathbf{Q}_h$ is the normalization of data description in the image, where $\mathbf{P}_3$ is the third column of the camera model **P**. This allows to fix the scale factor and to define the error vector: $[\varepsilon] = \begin{pmatrix} [\varepsilon_1] & [\varepsilon_2] & 0 \end{pmatrix}^t$. According to the model, there is no error on the scale factor, but only an uncertainty on the position of the geometric point in the image plane. $[\mathbf{q}]$ is the resulting interval vector. Values of $[\mathbf{q}]$ describe the boundaries of the projections of the 3D point in the image plane. The pixels position are described with intervals $([\mathbf{q}_l], [\mathbf{q}_r])$. From [14] it provides the system (equation 7) based on interval arithmetic rules. First, the matrix **P** associated to a camera model is cut such as:

$$\mathbf{P} = (\mathbf{M} \mid \mathbf{V}) \tag{6}$$

Where **M** is a (3×3) matrix and **V** is a (3×1) vector. From equation 6, and by introducing the operator $[*]_\times$ then the system to solve (equation 7) may be written in the interval analysis framework as:

$$[\mathbf{A}]\,\mathbf{Q}_{nh} = [\mathbf{B}] \tag{7}$$

with

$$[\mathbf{A}] = \begin{pmatrix} [[\mathbf{q}_l]]_\times \mathbf{M}_l \\ [[\mathbf{q}_r]]_\times \mathbf{M}_r \end{pmatrix} ; [\mathbf{B}] = \begin{pmatrix} [[\mathbf{q}_l]]_\times \mathbf{V}_l \\ [[\mathbf{q}_r]]_\times \mathbf{V}_r \end{pmatrix} \tag{8}$$

where $[\mathbf{A}]$ is an interval matrix, $[\mathbf{B}]$ an interval vector, and $[*]_\times$ the cross product function. This operator gives the associate anti-symmetrical matrix. For a given interval vector this operator is such as: $([a][b][c])^t$

$$\left[\begin{pmatrix} [a] \\ [b] \\ [c] \end{pmatrix}\right]_\times \mapsto \begin{pmatrix} 0 & [-c] & [b] \\ [c] & 0 & [-a] \\ [-b] & [a] & 0 \end{pmatrix}$$

The exact set of 3D points $\{\mathbf{Q}_s\}$ which is solution of the uncertain linear system is :

$$\{\mathbf{Q}_s\} = \left\{\mathbf{Q}_{nh} \in \mathbb{R}^3 \mid \exists \mathbf{A} \in [\mathbf{A}], \exists \mathbf{B} \in [\mathbf{B}], \mathbf{A}\mathbf{Q}_{nh} = \mathbf{B}\right\} \tag{9}$$

In the framework of interval analysis, linear system such as equation 9 can be solved using a *fixed point contractor* [15]. The use of this tool in computer vision has been developed in [1]. Applied to the linear system given by equation 9 it provides a box $[\mathbf{Q}_s]$ which contains the solution set $\{\mathbf{Q}_s\}$ such as:

$$[\mathbf{Q}_s] \quad = [\{\mathbf{Q}_{nh} | \exists \mathbf{A} \in [\mathbf{A}], \exists \mathbf{B} \in [\mathbf{B}], \mathbf{A}\mathbf{Q}_{nh} = \mathbf{B}\}] \tag{10}$$

Let's call $\mathcal{C}_{GS}$ the Gauss-Siedel contractor and $\mathcal{C}_K$ the Krawczyk contractor. Both seek for the minimal $[X_s]$ such as:

$$\begin{array}{ll} \{\mathbf{Q}_s\} & \subset [\mathbf{Q}_s] = \mathcal{C}_{GS}([\mathbf{A}], [\mathbf{B}]) \\ \{\mathbf{Q}_s\} & \subset [\mathbf{Q}_s] = \mathcal{C}_K([\mathbf{A}], [\mathbf{B}]) \end{array} \tag{11}$$

Applying these operators solve the uncertain linear system 8 for a couple of calibrated camera and a set of matched points. In [1] a comparison is given which led us to choose the Gauss Siedel contractor as it provides a good trade-off between accuracy and speed. Figures 1 give some examples of depth map computation using Interval Analysis.

Interestingly, this is an inconvenient of Interval Analysis which insures us that locally connected points will be merged. Indeed the main problem related to bounding box representation of space is the wrapping effect. More precisely the box provided is aligned with the reference frame and might not give a good approximation of the true shape of the space where the 3D point might lie. The side effect is that bounding boxes of nearby points intersect. In this paper this default is used to merge the points.

```
Algorithm:3D region growing using Interval Analysis
Data: $\mathcal{D}$
Result: List of possible objects $O$
$O = \emptyset$;
for $i \leftarrow 1$ to $|D|$ do
    $[\mathbf{x}] = D[i]$
    Merged $\leftarrow$ false
    Exploration $\leftarrow$ true
    $j \leftarrow 0$
    while Exploration do
        if $j < |O|$ then
            if $[x] \cap OuterBoundingBox(o)$ then
                Connected $\rightarrow$ false
                $Exploration_o \rightarrow$ true
                $k \leftarrow 0$
                while $Exploration_o$ do
                    $[\mathbf{y}] \leftarrow o[k]$
                    if $[x] \cap [y]$ then
                        Merged $\leftarrow$ true
                    end
                    $k \leftarrow k + 1$
                end
                if Merged = false then
                    $o \leftarrow o \cup [\mathbf{x}]$
                    $OuterBoundingBox(o) \leftarrow$
                    $Max(OuterBoundingBox(o), [\mathbf{x}])$
                    $InnerBoundingBox(o) \leftarrow$
                    $Max(InnerBoundingBox(o), \mathbf{x})$
                    Exploration $\leftarrow$ false
                end
            end
        else
            Exploration $\leftarrow$ false
        end
        $j \leftarrow j + 1$
    end
    if Merged = false then
        Create a new region $o$
        $o \leftarrow [\mathbf{x}]$
        $O \leftarrow O \cup o$
    end
end
```

VI. 3D REGION GROWING USING UNCERTAIN POINTS

The algorithm for 3D region growing using uncertain points takes as an input a dense map named D, and output a list of objects. The dense map provided by the vision system is given according to the image topology. However due to the iso-luminance filtering some points might be removed. A point of the map is noted $[\mathbf{x}]$. The algorithm maintain a list of regions named O which are coded as bounding boxes. The point is tested againt each region of

O. Each region has two bounding boxes : one based on the centre of each point (the *Inner* bounding box), the other based on the bounding box of each points belonging to the region (the *Outer* bounding box). Each point on the range map is tested across the existing regions, if the bounding box of a region intersect with the interval of a 3D point then this point is a potential candidate for merging.

Once a potential region has been found, the candidate bounding box should intersect at least the bounding box of another point, like depicted in figure 2-(b). Otherwise, the candidate is in the case depicted in figure 2-(a), and can not be merged.

If a point is merged, the outer bounding box of the region is updated by testing if the limits of the interval provided by $[\mathbf{x}]$ expand its own limits. The inner bouding box is updated by considering only the center of $[\mathbf{x}]$ which we note $\mathbf{x}$. Finally if the point is left alone it creates a new box.

As the point are tested following the image topology, the last points merged are put at the beginning of the region's list.

In the pre-attentive behaviour the target is chosen as the region with the highest number of points. Figure 3-f shows the result of the segmentation on the scene represented in figure 3 (a-e). The blue box is the inner bounding box, while the red box is the outer bounding box.

VII. Experiment

A. Context

The experiments are realized on a humanoid robot HRP-2 [11]. In the head of this robot, 4 cameras are embedded. Three are used for 3D model-based object recognition [16]. They are rigidly fixed to the head, and then might be precisely calibrated. The fourth one has a wide field of view for visual feedback during teleoperation. In this paper only two have been used. Also this robot has two Pentium PIII 1 GHz CPU boards, only one is used to perform the computation related to vision. The software structure of this system relies on CORBA to add incrementally modules, and CPUs. A specific architecture exists concerning real-time issues for controlling the robot. This architecture is described more precisely in [17].

The disparity is computed using a modified version of the VVV software presented in [16]. This software has been reorganized to offer a flexible interface for higher level processes. It is possible to start or stop on-line visual processes, and change their parameters. Using CORBA, it is possible to control efficiently the processes, and get the result in various languages and platform.

The robot is placed 2 meters far away from a table on top of which is a cookie box. Using the algorithm described in section V, a dense map is computed. It is used as the input of the algorithm described in section VI. The regions of interest are sorted according to their number of points. In this particular case, the table does not have texture, and therefore almost all the points of its upper part are discarded. The floor is also suppressed, and consequently the cookie box is picked up as the main point of interest. Once the position of the target has been found into the vision system reference frame, it is projected back into the world reference frame of the robot. Finally this information is send to the pattern generator to put the robot 50 cm before the object of interest. The full sequence is depicted in figure 3.

B. Discussion

Also it has been possible to successfully implement this pre-attentive behaviour they are several limitations. The first limitation is due to the distance used. As it is purely geometrical they are no difference between the object and its immediate surrounding, For instance the cookie box and the edge of the table are merged together in figure 1-f. Moreover, it assumes sufficient texture to have enough 3D points. Those are classical drawbacks in such technique. They are at least two solutions to fix the problem: one is to use the pixel intensity as a supplementary information as in [12], the second is to make the robot interacts with the object for further refinement. Both solutions are currently under investigation. The second solution has the advantage of integrating haptic information.

The second current limitation is due to the implementation of this solution. The validation of this algorithm has been realized by solving the system given by equation 7 for each matching point. In [1] this cost has been measured to be 5*ms*. As the map used in figure 3 contains 40,000 points, it takes 3 minutes to be computed. A possible solution could be to build an approximation function

$$\hat{f}(\mathbf{q}_r, \mathbf{q}_l) = [\mathbf{Q}_s] \tag{12}$$

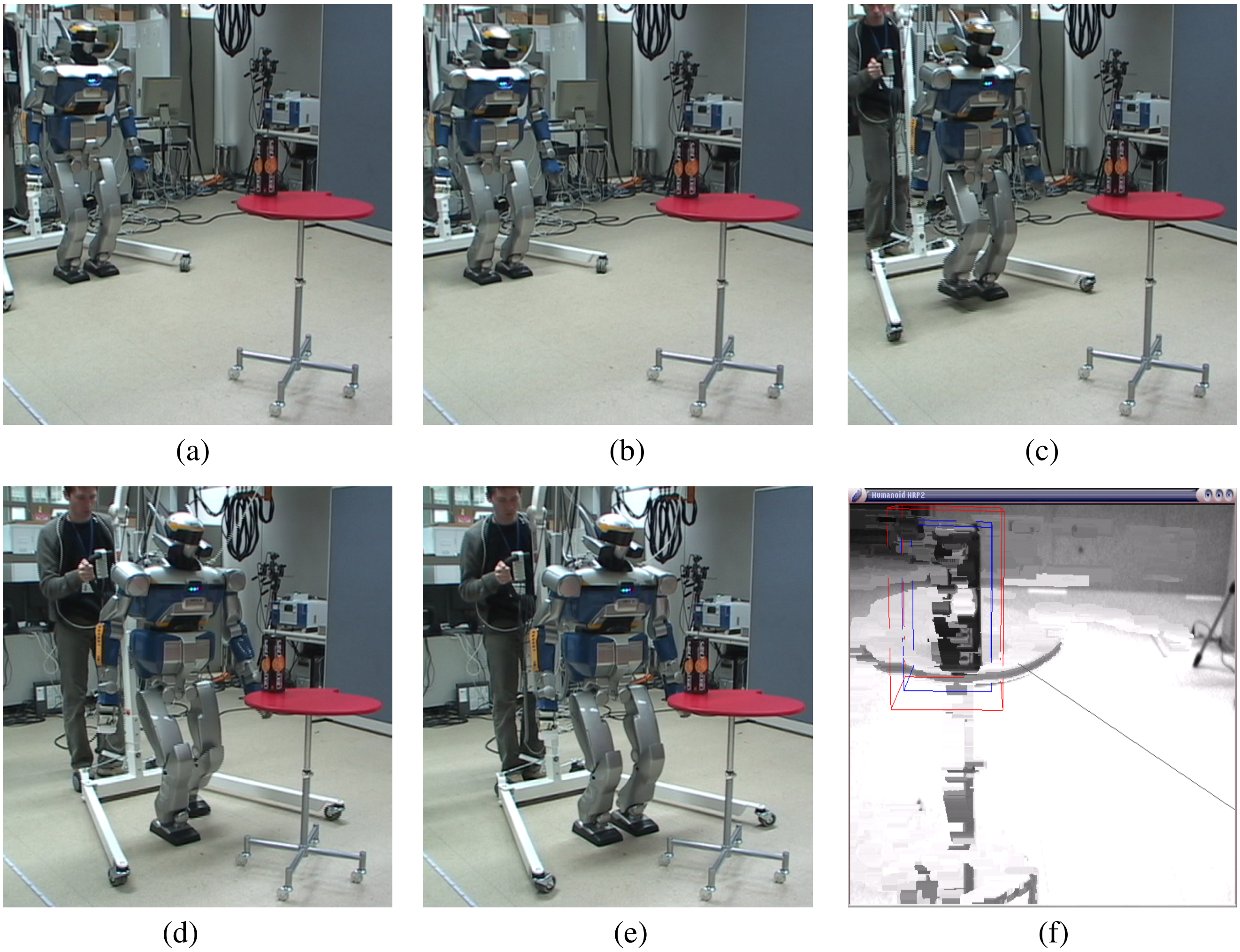
The main problem will be to insure that this approximation keeps the upper bound property provided by the interval analysis framework.

What could be a third limitation is the estimation of ε. Indeed if the threshold has disappeared from the segmentation algorithm, a new parameter has been introduced into the camera model. However in the experiment described here this parameter was set to 0.5 which is equivalent to the pure geometrical error reconstruction. More generally, this parameter depends onto the matching process error. Thus it is not a new parameter of the segmentation itself.

VIII. Conclusion

We have proposed a 3D segmentation algorithm using the 3D reconstruction error estimation provided by the Interval Analysis framework. This allows us to not use any threshold for merging points. It has been implemented and used for realizing a pre-attentive behaviour where a humanoid robot goes towards an unknown object. The main advantage of this approach is to rely mainly onto the intrinsic parameters of the robot, here the ones related to its vision system.

The authors would like to thanks the Japanese Society for Promotion of Science for partial funding of this work.



(a) (b) (c)

(d) (e) (f)

Fig. 3

THE HRP-2 HUMANOID ROBOT STOPPING 50 CM BEFORE THE OBSTACLE (A-E) AFTER DETECTION USING INTERVAL ANALYSIS (F).

REFERENCES

[1] B. Telle, O. Stasse, T. Ueshiba, K. Yokoi, and F. Tomita, “3d boundaries partial representation of objects using interval analysis,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, November 2004, pp. 4013–4018.

[2] O. Faugeras, *Three Dimensional Computer Vision*. MIT press, 1992.

[3] G. Gelade and A. Treisman, “A feature-integration theory of attention,” *Cognitive psychology (1980)*, vol. 12, pp. 97–136, 1980.

[4] J. A. Driscoll, R. A. P. II, and K. S. Cave, “A visual attention network for a humanoid robot,” in *International Conference on Intelligent Robotic Systems*, October 1998, pp. 1968–1974.

[5] Olivier Stasse, Yasuo Kuniyoshi, Gordon Cheng, “Development of a Biologically Inspired Real-Time Visual Attention System,” in *Biologically Motivated Computer Vision, LNCS 1811*, Seoul, Korea, 2000, pp. 150–159.

[6] T. Minato and M. Asada, “Towards selective attention: generating image features by learning a visuo-motor map,” *Robotics and Autonomous Systems*, vol. 45, pp. 211–221, 2003.

[7] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature*, vol. 2, pp. 194–203, March 2001.

[8] J. K. Yi Ma, Stefano Soatto and S. S. Sastry, *An invitation to 3-D Vision*, I. S. S.S. Antman, J.E. Marsden and S. Wiggins, Eds. Springer-Verlag, Interdisciplinary Applied Mathematics, Imaging, Vision and Graphics, 2004.

[9] S. Lazebnik, C. Schmid, and J. Ponce, “Semi-local affine parts for object recognition,” in *British Machine Vision Conference*, September 2004, pp. 959–968.

[10] K. Yamazaki, M. Tomono, T. Tsubouchi, and S. Yuta, “Object shape reconstruction and pose estimation by a camera mounted on a mobile robot,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 4019–4025.

[11] K.Kaneko, F.Kanehiro, S.Kajita, H.Hirukawa, T.Kawasaki, M.Hirata, K.Akachi, and T.Isozumi, “Humanoid robot hrp-2,” in *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, vol. 2, 2004, pp. 1083–1090.

[12] Z. Lin, J. Jin, and H. Talbot, “Unseeded region growing for 3d image segmentation,” in *ACM International Conference Proceeding Series, Selected papers from the Pan-Sydney workshop on Visualisation*, vol. 2, 2000, pp. 31–37.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry*. Cambridge University Press, 2003, major book on 3D vision.

[14] B. Telle, “Méthode ensembliste pour une reconstruction 3d garantie par stereo vision,” Ph.D. dissertation, Université Montpellier II, 2003.

[15] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis*. London: Springer Verlag, 2001.

[16] Y. Sumi, Y. Kawai, T.Yoshimi, and T. Tomita, “3d object recognition in cluttered environments by segment-based stereo vision,” *International Journal of Computer Vision*, vol. 6, pp. 5–23, January 2002.

[17] F. Kanehiro, H. Hirukawa, and S. Kajita, “Openhrp: Open architecture humanoid robotics platform,” *The International Journal of Robotics Research*, vol. 23, no. 2, pp. 155–165, 2004.