Stereo Vision for Obstacle Detection: A Region-Based Approach

P. Foggia[2], A. Limongiello[1] and M. Vento[1]

[1] Dip. di Ingegneria dell'Informazione ed Ingegneria Elettrica
Università di Salerno, Via Ponte don Melillo, I84084 Fisciano (SA), Italy

[2] Dip. di Informatica e Sistemistica
Università di Napoli, Via Claudio 21, I80125, Napoli, Italy

Abstract. We propose a new approach to stereo matching for obstacle detection in the autonomous navigation framework. An accurate but slow reconstruction of the 3D scene is not needed; rather, it is more important to have a fast localization of the obstacles to avoid them. All the methods in the literature, based on a pixel stereo matching, are ineffective in realistic contexts because they are either computationally too expensive, or unable to deal with the presence of uniform patterns, or of perturbations between the left and right images. Our idea is to face the stereo matching problem as a matching between homologous regions. Our method is strongly robust in a realistic environment, requires little parameter tuning, and is adequately fast, as experimentally demonstrated in a comparison with the best algorithms in the literature.

1 Introduction

During the last years, the Computer Vision community has shown an increasing interest in applications like Automated Guided Vehicles (AGV) or Autonomous Mobile Robots (AMR). In the literature many approaches have been proposed for Visual Navigation of a mobile platform [1],[2]. A very challenging task is the so-called *obstacle detection*. Many authors have expressed their conviction that a robotic vision system should aim at reproducing the human vision system, and so should be based on stereo vision. The greatest advantage of stereo vision with respect to other techniques (e.g. optical flow, or model-based) is that it produces a full description of the scene, can detect motionless and moving obstacles (without defining a complex obstacle model), and is less sensitive to the environmental changes (the major disadvantage of optical-flow techniques). The stereo vision provides a "3D-like" representation of the scene, producing information about objects in the environment that may obstacle the motion. A pair of images acquired from a stereo camera implicitly contains depth information about the scene: this is the main assumption of stereo vision. The main difficulty is to establish a correspondence between points of the two images representing the same point of the scene; this process is called *disparity matching*.

Foggia P., Limongiello A. and Vento M. (2007).
Stereo Vision for Obstacle Detection: A Region-Based Approach.
In *Robot Vision*, pages 36-45
DOI: 10.5220/0002067900360045

The set of displacements between matched pixels is usually indicated as *disparity map*. All the approaches, in the literature, are based on this pixel correspondence. We propose an extension of that concept, namely we define a disparity value for a whole region of the scene starting from the two homologous views of it in the stereo pair. The main reason of this extension is that a pixel-matching approach is redundant for AMR and AVG applications. In fact, in this framework, it is not very important to have a good reconstruction of the surfaces, but it is more important to identify adequately the space occupied by each object in the scene, even by just assigning to it a single disparity information. Moreover the pixel-based approaches are lacking in robustness in some realistic frameworks, especially for video acquired from a mobile platform. Our method estimates the average depth of the whole region by an integral measure, and so has fewer problems with uniform regions than other methods have. The estimate of the position of the regions is sufficiently accurate for navigation and it is fast enough for real time processing.
This paper is organized as follows: Section 2 presents the state of art for stereo matching problem; Section 3 is devoted to show the rationale of our method; Section 4 shows the algorithm. Finally, in Section 5 there is a discussion of experimental results on a standard stereo image database and also on our stereo video sequences. Conclusions are drawn in Section 6.

2 Related Works

We will present a brief description of the most important methods for stereo matching; for more details, there is a good taxonomy proposed by Scharstein and Szeliski [3], and a survey on stereo vision for mobile robots by Zhang [4]. There are two major types of techniques, in the literature, for disparity matching: the area-based and feature-based techniques. Moreover, the area-based algorithms can be classified in local and global approaches. The local area-based algorithms [5],[6],[7] provide a correspondence for each pixel of the stereo pair. They assume that each pixel is surrounded by a window of pixels having similar disparity; these windows are matched using correlation or a similar technique. They produce a dense disparity map (i.e. a map providing a disparity for each pixel), more detailed than it is needed for AMR aims. Furthermore, they can be quite unreliable, not only in homogeneous regions, but also in textured regions for an inappropriately chosen window size. On the other side, the global area-based approaches (that also yield a dense map) try to propagate disparity information from a pixel to its neighbours [8],[9], or they define and minimize some energy function over the whole disparity map [10],[11],[12]. They have a better performance in homogeneous regions, but they frequently have parameters which are difficult to set, and are highly time-consuming. The feature-based approaches [13],[14],[15] detect and match only "feature" pixels (as corner, edges, etc.). These methods produce accurate and efficient results, but compute sparse disparity maps (disparity is available only in correspondence to the feature points). AMR applications require more details, such as some information about the size of the objects; also a rough shape of the objects is needed for guiding a robot in the environment or for basic recognition tasks (e.g. in industrial applications, or for platooning of

robots). All the proposed methods, as already said, look for a pixel matching in the stereo pair. Therefore, some constraints have been introduced, since the first works on the stereopsis by Marr and Poggio [8],[13] in order to guarantee good results. The stereo pair is supposed to be acquired from a sophisticated system, so that the intensity distributions of the two images are as similar as possible. Moreover, a pre-processing phase is needed, before the correspondence finding step, to compensate the hardware setup (*calibration* phase), or to assume an horizontal epipolar line (*epipolar rectification*). Unfortunately, in realistic applications of mobile platforms these constraints are not easy to guarantee. The two images of the stereo pair could have a different intensity distribution, the motion of the mobile platform on a rough ground could produce mechanical vibrations of the cameras, and consequently local or global perturbations between the two images.

3 The Rationale

In this paper we propose an extension of the disparity concept. The main idea is to determinate a unique disparity value for a whole region of the scene and not for a pixel. In fact, even if we can suppose a unique correspondence between each pixel in the left and right images from an optical point of view, in some cases we can not have enough information to find this correspondence looking just at a single pixel. Let us consider three kinds of situations:

Pixels inside homogeneous areas. The features-based algorithms are unable to find an appropriate feature in case of textureless region. The local area-based techniques must define a big correlation area in order to pick enough information for the matching. Finally, the global area-based methods produce a propagation of the error depending on the energy minimization.

Local and global perturbation of the stereo pair depending on the vibration of the mobile platform. The motion of the robot produces mechanical vibrations of the cameras with a consequent loss of the horizontal epipolar line constraint, which is assumed from all the methods in the literature.

A different intensity distribution between the left and right image. In a realistic framework the stereo pair could suffer from perspective or photometric distortions. Moreover, the two cameras could have different acquiring parameters, i.e. focus, or exposure, etc. In ideal conditions all the pixels belonging to the same depth level have two intensity patterns between the left and right image with a unique horizontal displacement. In real condition the two intensity patterns are no longer a simple horizontal translation, consequently a pixel matching could be unsuitable.

A region-based algorithm is proposed to face up the limitations of the pixel stereo matching approaches. The corresponding entity is no longer the pixel, but a region; the matching of regions provide a lowering of resolution, but an increasing of robustness in a realistic environment. In fact, a uniform area is considered as a unique segment for the matching, so that the local and global perturbations of the stereo pair less influence the solution. Finally, the proposed algorithm (see next section) define a matching function, which is able to mitigate the lack of homogeneity between the left

and right image. Therefore, a good tread-off, between an efficient solution (to guarantee an autonomous navigation) and the robustness in a realistic framework, is investigated. Moreover, the real-time requirement is guaranteed.

4 The Algorithm

We determinate the disparity value for the whole region as the horizontal displacement between the regions. The detection of the homologous regions is, of course, a difficult problem. In fact, a same segmentation method, separately applied on the left and right image, should divide in different parts the same region of the scene, or should produce border errors, undermining a correct detection of the disparity for the whole region. In our algorithm a segmentation of the left image (reference image) is performed and each segment is overlapped on the sensed image (right image). The disparity value of the region is the horizontal displacement, corresponding to the minimization of a *best fitting function* between the two regions.

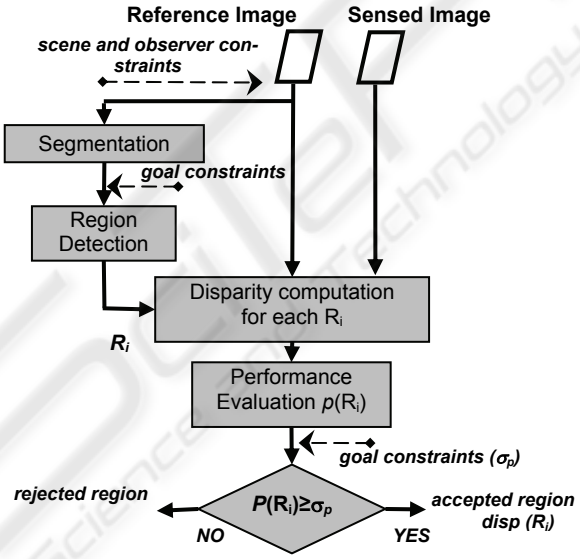

Fig. 1. A schema for our algorithm of region-based stereo matching.

This integral measurement of the disparity can mitigate some null integral border errors, as segmentation, digitalization, and photometric errors. An approximation can be obtained for the border errors from perspective distortion, that is not right with null integral. It should be noted that a region is not an object; objects are decomposed into several regions, so the overall shape of the object is however reconstructed. As shown in Fig. 1 the algorithm is composed of four steps:

Segmentation of the reference image. Several segmentation methods (mean shift, pyramid, multi-threshold) have been tested. The algorithm has a similar behavior towards all the methods, taking care not to under segment the image. In fact, an under

segmentation could merge regions belonging to different depth level. The over segmentation has not a big influence in our method, because the best fitting function is enough accurate.

Region Detection. A connected component analysis is performed to detect connected segments. Looking at the experimental results, a 4-connected analysis has been enough for our aim. This step is also devoted to select a subset of regions among all. The selection is made using some constraints on the goal (*goal constraints*). Namely, a minimal knowledge about the obstacle (i.e. the maximum size of an obstacle is an upper-bound for the maximum size of a region; color information if any) or the desired resolution of the result. In this way, the computation time can be reduced.

Disparity Computation. Each segment from Region Detection step is used as a selection mask on the left and right image in order to select the homologous regions. The selection of the right region is displaced from 0 to the maximum value of disparity. The disparity is the horizontal displacement corresponding to the best fitting of the homologous regions. Formally, let $E_L(x,y)$ and $E_R(x,y)$ be the intensity value for each pixel (x,y) on the left and right image. Let $G_L(x,y)$ and $G_R(x,y)$ be the gradient map of the left and right images. Finally, let R_i be the generic segment from step 2, the following equations are defined:

$$d(R_i) = \arg\min_{0 \le d \le d_{\max}} (\varepsilon_i(d)) \qquad \varepsilon_i(d) = \alpha \cdot \varepsilon_i^{col}(d) + \beta \cdot \varepsilon_i^{grad}(d) \tag{1}$$

The best fitting function uses color and gradient information in order to consider the intensity distribution of pixels inside each region and also texture information. The values for the weights α and β are experimentally found.

$$\varepsilon_i^{col}(d) = \frac{1}{|R_i|} \sum_{(x,y) \in R_i} \left| \left(E_L(x,y) - \mu_i^L\right) - \left(E_R(x-d,y) - \mu_i^R(d)\right) \right| \tag{2}$$

$$\mu_i^L = \frac{1}{|R_i|} \sum_{(x,y) \in R_i} E_L(x,y) \qquad \mu_i^R(d) = \frac{1}{|R_i|} \sum_{(x,y) \in R_i} E_R(x-d,y)$$

$$\varepsilon_i^{grad}(d) = \frac{1}{|R_i|} \sum_{(x,y) \in R_i} |G_L(x,y) - G_R(x-d,y)|$$

The best fitting color function, $\varepsilon_i^{col}(d)$, is normalized on the mean color (μ_i^L , μ_i^R) of the region R_i on the left and right image. In this way a good matching is found also in case of a not homogeneous distribution of intensity between the left and right image.

Performance Evaluation. The previous step provides also a performance index for the matching, $p(R_i)$. For each region the minimum value of the fitting function has been used as matching error, $\varepsilon(R_i)$, and the reliability index is:

$$p(R_i) = 1 - \frac{\varepsilon(R_i)}{\max_{R_i}(p(R_i))} \qquad where: \qquad \varepsilon(R_i) = \min_{0 \le d \le d_{\max}} (\varepsilon_i(d)) \tag{3}$$

A region is rejected if the performance index is lower than an imposed tolerance

($p(R)<\sigma_p$). This is an other goal constraint because we can choose a reliability level depending on the requested efficacy of the solution.

Vision is an under constrained problem and therefore that it is necessary to find new constraints in order to make the problem solvable. Our method can be classified as a systemic approach [17], in fact we consider constraints coming from the scene, from the goal and from the observer (as shown in Fig. 1). In particular, with regard to scene constraints, we assume a strong continuity constraint for each selected region, and the compatibility and the uniqueness constraints are applied on the whole region and not longer on each pixel. Moreover, the observer (mobile platform) moves slowly so that only little vibrations of the cameras are possible. Finally, the obstacle detection task (our goal) is scalable in time and performance: a robot, having more time, can carry out a finer investigation of the environment, asking to the system a better solution.

5 Experimental Results

In the literature, tests are usually performed with standard databases composed of static images, well-calibrated and acquired in uniform lighting. The Middlebury web site by Scharstein and Szeliski [18] is a good reference for some stereo images and to compare some stereovision algorithms. In this section we want to show our qualitative results and discuss some errors of the best algorithms in the literature, when applied to real cases. Nowadays, in AMR and AGV applications it is not defined a quantitative measurement for performance evaluation. In [16] it is proposed a quantitative performance evaluation for disparity map, but in case of reconstruction aims. For this reason we also propose a quantitative method to compare stereo algorithms when the goal is the obstacle detection and no longer the 3D reconstruction of the scene. The following Fig. 2 shows our result on the Tsukuba DB and a comparison with other approaches. We have selected: squared differences (SSD) and graph cuts (GC) [18] (a local and global area-based algorithm respectively). The experiments have been performed on a notebook Intel P4 1.5 GHz, 512 Mb RAM, and we have considered a resolution of 384x288 pixel. We have used the following parameters and constraints:

Table 1. Parameters and constraints.

Description	Value
Scene and observer constraints	Respected
Numbers of ranges for segmentation	20
Goal constraint for region detection	Not used
$[\alpha , \beta]$ for disparity computation	[0.4, 0.6]
Threshold for performance evaluation (σ_p)	0.8

The goal constraint for region detection is not used in order to not compromise the comparison with the pixel-based approaches (that can't use such constraint). The parameters are obtained by experimental evidences.

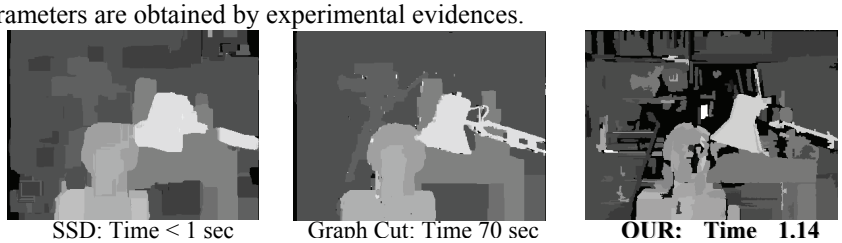

Fig. 2. A comparison with other approaches.

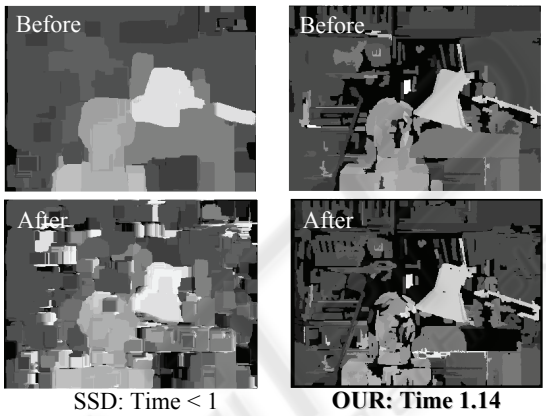

Fig. 3. SSD and Our approach after a vertical translation of 2 pixels.

In Fig. 3 it is clear the robustness of our approach in relation to the loss of the horizontal epipolar constraint.

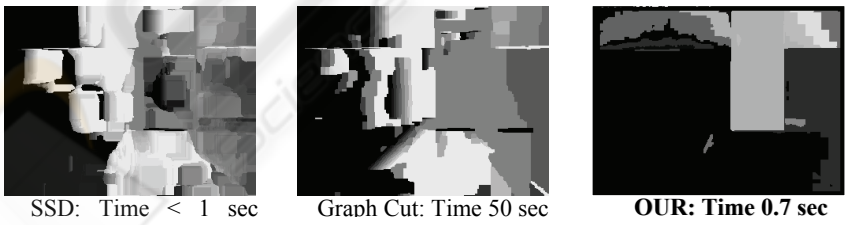

Fig. 4. Results on our stereo pair: it is characterized by only one homogeneous object.

The presence of texture-less regions causes serious problems to the best algorithms of the literature as shown in Fig. 4. In order to consider a quantitative comparison of the algorithms for obstacle detection aim, we define a simple module that detects the obstacles from the disparity map. Each 4-connected region with the same disparity value is identified with a bounding box and its distance from the observer. We select the obstacles as the connected regions that belong to a chosen range of distances, in

fact an obstacle is an object so close to the mobile platform to forbid the navigation. Two performance index are defined in order to valuate: the capability of the algorithm to identify adequately the space occupied by each obstacle (*occupancy performance*); the correctness of depth computation for each obstacle (*distance performance*). For each frame of the video sequence acquired from the platform, let R_G be the real obstacle regions (*Ground Truth*), let R_D be the obstacle regions detected by the algorithm, and let R_I be the subset of regions correctly detected as obstacles by the algorithm ($R_I = R_G \cap R_D$). The occupancy performance is evaluated with the measures of *precision* and *recall*:

$$recall = \frac{R_I}{R_G} \qquad precision = \frac{R_I}{R_D} \tag{4}$$

The distance performance is evaluated with a *relative distance error (rde)*:

$$rde = \frac{|\text{detected distance} - \text{real distance}|}{\text{real distance}} \tag{5}$$

The distance of an obstacle is related to its disparity value following the relation:

$$\text{distance} = k_{px/m} \frac{baseline \cdot focal\ lenght}{disparity} \tag{6}$$

where $k_{px/m}$ is the conversion factor from pixel to meter. It should be noted that for each real obstacle (*Ground Truth*) could be more than one overlapped obstacle regions detected by the algorithm. The detected distance is supposed to be a weighted mean distance of all the overlapped regions. The weights are set up to the sizes of each overlapping area. We report some results obtained on a realistic video acquired from our mobile platform. The video sequence (100 frames) is characterized by camera vibration, light changing, uniform obstacles (see Fig. 5).

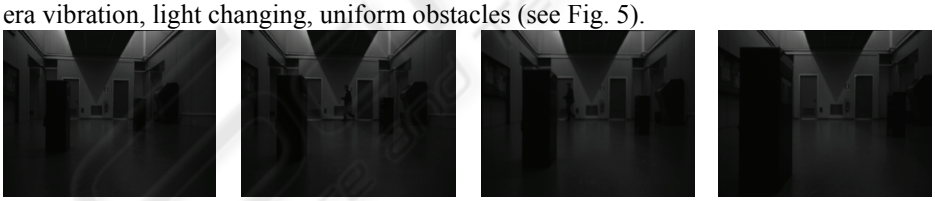

Fig. 5. Some frames of the video sequence.

The proposed method is compared with the Small Vision System (SVS) by Konolige [20,21], that is the most popular system in off-the-shelf systems. We consider two different version of that algorithm: *SSD* and *SSD multi-scale*.

Fig. 6. Disparity Map Results: On the left side our method, on the center side the SSD, and on the right side the SSD multi-scale.

Table 2. Precision and Recall.

algorithm	recall	precision
our method	**0.910500462**	**0.635458231**
SSD	0.208928846	0.477767
SSD multi-scale	0.469375385	0.348905692

Table 3. Relative Distance Error.

algorithm	relative distance error
our method	**0.108793103**
SSD	0.191169769
SSD multi-scale	0.180161282

The results in the previous tables show that our method is much better than the other two, especially for the occupancy performance. In fact, as it is clear from Fig. 7 and Fig. 8, our approach can better overlap the space occupied by the real obstacles, as like the SSD multi-scale algorithm has big opening areas inside the obstacles.

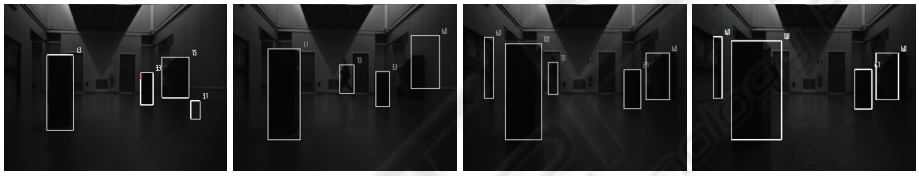

Fig. 7. Some results of our obstacle detection algorithm.

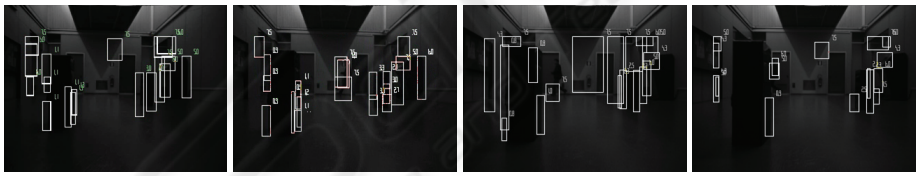

Fig. 8. Some results of obstacle detection from SSD multi-scale stereo algorithm.

6 Conclusions

We have presented a stereo matching algorithm that is especially oriented towards AMR and AGV applications, providing a fast and robust detection of object positions instead of a detailed but slow reconstruction of the 3D scene. The algorithm has been experimentally validated showing an encouraging performance when compared to the most commonly used matching algorithms, especially on real-world images. The bigger problem is uniform areas, and here clearly correlation-based stereo does not work, giving only the edges of the regions. Future works are oriented to combine region-matching with conventional correlation stereo and to develop a temporal coherence of the solution in the video sequence.

References

1. DeSouza, G. N. , Kak, A. C.: Vision for Mobile Robot Navigation: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 237-267 (2002).
2. Kastrinaki, V., Zervakis, M., Kalaitzakis, K.: A survey of video processing techniques for traffic applications. *Image and Vision Computing*, vol. 21, pp. 359–381 (2003).
3. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision, vol. 47, no. 1, pp. 7-42 (2002).
4. Zhang, C.: A Survey on Stereo Vision for Mobile Robots. Technical report, Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA (2002).
5. Kanade, T., Okutomi, M.: A stereo matching algorithm with an adaptive window: Theory and experiment. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 16, no. 9, pp. 920-932 (1994).
6. Fusiello, A., Roberto, V.: Efficient stereo with multiple windowing. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 858–863, Puerto Rico (1997).
7. Veksler, O.: Stereo matching by compact windows via minimum ratio cycle. Proceedings of the International Conference on Computer Vision, vol. I, pp. 540–547, Vancouver, Canada (2001).
8. Marr, D., Poggio, T.A.: Cooperative computation of stereo disparity. Science, vol. 194, no. 4262, pp. 283–287 (1976).
9. Zitnick, C.L., Kanade, T.: A cooperative algorithm for stereo matching and occlusion detection. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22, no. 7, pp. 675–684 (2000).
10. Geiger, D., Ladendorf, B., Yuille, A.: Occlusions and binocular stereo. International Journal of Computer Vision, vol. 14, pp. 211–226 (1995).
11. Roy, S.: Stereo without epipolar lines: A maximum-flow formulation. International Journal of Computer Vision, vol. 34, no. 2/3, pp. 1–15 (1999).
12. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222–1239 (2001).
13. Marr, D., Poggio, T.A.: A computational theory of human stereo vision. RoyalP, vol. B, no. 204, pp. 301–328 (1979).
14. Grimson, W.E.L.: A computer implementation of a theory of human stereo vision. Royal, vol. B, no. 292, pp. 217–253 (1981).
15. Candocia, F., Adjouadi, M.: A similarity measure for stereo feature matching. IEEE Transaction on Image Processing, vol. 6, pp. 1460-1464 (1997).
16. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 195-202, Madison, WI (2003).
17. Jolion, J.M.: Computer Vision Methodologies. CVGIP: Image Understanding, vol. 59, no. 1, pp. 53–71 (1994).
18. http://cat.middlebury.edu/stereo/
19. http://www.videredesign.com/
20. Konolige, K.: Web site. http://www.ai.sri. com/software/SVS (2006).
21. Konolige, K.: Small vision systems: hardware and implementation. Intl. Symp. On Robotics Research, pages 111–116 (1997).