Obtaining Reliable Depth Maps for Robotic Applications from a Quad-Camera System

Lazaros Nalpantidis, Dimitrios Chrysostomou, and Antonios Gasteratos

Laboratory of Robotics and Automation
Department of Production and Management Engineering
Democritus University of Thrace
University Campus Kimmeria
GR-671 00, Xanthi, Greece
{lanalpa,dchrisos,agaster}@pme.duth.gr
http://robotics.pme.duth.gr

Abstract. Autonomous navigation behaviors in robotics often require reliable depth maps. The use of vision sensors is the most popular choice in such tasks. On the other hand, accurate vision-based depth computing methods suffer from long execution times. This paper proposes a novel quad-camera based system able to calculate fast and accurately a single depth map of a scenery. The four cameras are placed on the corners of a square. Thus, three, differently oriented, stereo pairs result when considering a single reference image (namely an horizontal, a vertical and a diagonal pair). The proposed system utilizes a custom tailored, simple, rapidly executed stereo correspondence algorithm applied to each stereo pair. This way, the computational load is kept within reasonable limits. A reliability measure is used in order to validate each point of the resulting disparity maps. Finally, the three disparity maps are fused together according to their reliabilities. The maximum reliability is chosen for every pixel. The final output of the proposed system is a highly reliable depth map which can be used for higher level robotic behaviors.

Keywords: Stereo vision, quad-camera system, disparity maps fusion.

1 Introduction

Reliable depth estimation is commonly needed in robotics in order to obtain numerous autonomous behaviors. Autonomous navigation [1], localization and mapping are just a few of them [2,3]. Vision-based solutions are becoming more and more attractive due to their decreasing cost as well as their inherent coherence with human imposed mechanisms. However, robotic applications place strict requirements on the demanded speed and accuracy of vision depth-computing algorithms. Depth estimation using stereo vision, comprises the stereo correspondence problem. Stereo correspondence is known to be very computational demanding. The computation of dense and accurate depth images, i.e. disparity maps, in frame rates suitable for robotic applications is an open problem for

M. Xie et al. (Eds.): ICIRA 2009, LNAI 5928, pp. 906–916, 2009.

the scientific community. Most of the attempts to confront the demand for accuracy focus on the development of sophisticated stereo correspondence algorithms, which usually increase the computational load exponentially. On the other hand, the need for real-time frame rates, inevitably, imposes compromises concerning the quality of the results. However, results' reliability is of crucial importance for autonomous robotic applications. This work proposes a quad-camera, depth estimation system able to produce reliable results while maintaining the computational load within acceptable limits.

1.1 Related Work

Autonomous robots behavior greatly depends on the accuracy of their decision making algorithms. In the case of stereo vision-based navigation, the accuracy and the refresh rate of the computed disparity maps are the cornerstone of its success [4]. Dense local stereo correspondence methods calculate depth for almost every pixel of the scenery, talking into consideration only a small neighborhood of pixels each time [5]. On the other hand, global methods are significantly more accurate but at the same time more computationally demanding, as they account for the whole image [6]. However, since the most urgent constraint in autonomous robotics is the real-time operation, such applications usually utilize local algorithms.

Muhlmann et al. in [7] describe a method that uses the sum of absolute differences (SAD) correlation measure for RGB color images. Applying a left to right consistency check, the uniqueness constraint and a median filter, it can achieve 20 fps for 160x120 pixel images. Another fast SAD based algorithm is presented in [8]. It is based on the uniqueness constraint and rejects previous matches as soon as better ones are detected. It achieves 39.59 fps speed for 320x240 pixel images with 16 disparity levels and the root mean square error for the standard Tsukuba pair is 5.77 . The algorithm reported in [9] achieves almost real-time performance. It is once more based on SAD but the correlation window size is adaptively chosen for each region of the picture. Apart from that, a left to right consistency check and a median filter are utilized. The algorithm is able to compute 7 fps for 320x240 pixel images with 32 disparity levels. A detailed taxonomy and presentation of dense stereo correspondence algorithms can be found in [5]. Additionally, the recent advances in the field as well as the aspect of hardware implementable stereo algorithms are covered in [10].

Early previous work focused on developing stereo algorithms mostly for binocular camera configurations. More recently, however, due to significant boost of the available computational power, vision systems using multiple cameras are becoming increasingly feasible and practical. The transition from binocular to multi-ocular systems has the advantage of potentially increasing the stability and accuracy of depth calculations. The continuous price-reduction of vision sensors allowed the development of multiple camera arrays ready for use in many applications. For instance, Yang et al. [11] used a five-camera system for real-time rendering using modern graphics hardware, while Schirmacher et al. [12] increased the number of cameras and built up a six-camera system for on-the-fly

processing of generalized Lumigraphs. Moreover, developers of camera arrays have expanded their systems so as to use tens of cameras, such as the MIT distributed light field camera [13] and the Stanford multi-camera array [14]. These systems are using 64, and 128 cameras respectively.

Most of the aforementioned camera arrays are utilized for real-time image rendering. On the other hand, a research area that could also be benefited by the use of multiple camera arrays is the so called cooperative stereo vision; i.e., multiple stereo pairs being considered to improve the overall depth estimation results. To this end, Zitnick [15] proposed an algorithm for binocular occlusion detection and Mingxiang [16] expanded it to trinocular stereo. The proposed system goes a step further using, for the first time, four cameras and exploiting the concept of the diagonal disparity.

2 Proposed System

The proposed system is a combination of sensory hardware and a custom-tailored software algorithm. The hardware configuration, i.e. the four cameras' formation, produce three stereo image pairs. Each pair is submitted to the simple and rapid stereo correspondence algorithm, resulting, thus, in a disparity map. For each disparity map a certainty map is calculated, indicating each pixel's reliability. Finally, the three disparity maps are fused, according to their certainties for each pixel. The outcome is a single disparity map which incorporates the best parts of its producing disparity maps. The combined hardware and software system is able to produce accurate dense depth maps in frame rate suitable for autonomous robotic applications.

2.1 Sensory Hardware Requirements

The sensory configuration of the proposed system consists of four identical cameras. The four cameras are placed so as their optical axes to have parallel orientation and their principal points to be co-planar, residing on the corners of the same square, as shown in Fig. 1(a). The images captured by the upper-left camera are considered as the reference images of each tetrad. Each one of the other three cameras produces images to be corresponded to the reference images. Thus, for each tetrad of images three, differently oriented, stereo pairs result, i.e. an horizontal, a vertical and a diagonal one. The concept, as well as the result of such a group of cameras are presented in Fig. 1(b).

2.2 Software Architecture

The proposed algorithm consists of two processing steps. The first one is the stereo correspondence algorithm that is applied to each image pair. Then, during a fusion step the results for all the stereo pairs are merged.

(a) (b)

Fig. 1. (a) The Quad-camera Configuration and (b) the Results (Up-Left) and Scene Capturing (Right) Using a Quad-camera Configuration

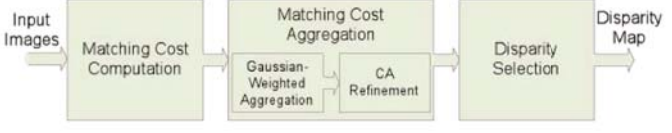

Fig. 2. Block Diagram of the Utilized Stereo Correspondence Algorithm

Stereo Correspondence Algorithm. The proposed system utilizes a custom tailored, simple, rapidly executed stereo correspondence algorithm applied to each stereo pair. Stereo disparity is computed using a three-stage local stereo correspondence algorithm. The algorithm utilized is an up-to-date version of the algorithm presented in [17]. It combines low computational complexity with sophisticated data processing. Consequently, it is able to produce dense disparity maps of good quality in frame rates suitable for robotic applications. The structural elements of this algorithm are presented in Fig. 2. The main attribute that differentiates this algorithm from the majority of the other ones is that the matching cost aggregation step consists of two sophisticated sub-steps rather than one simple summation. In addition, the disparity selection process is a non-iterative one and for speed reasons the final refinement step is absent. The results' refinement step is moved inside the aggregation, rather than being an additional final procedure. Instead of refining the results that were chosen through a strict selection process, the proposed algorithm performs a refinement procedure to all the available data. Such a procedure enhances the quality of the results. Thus, the disparity selection step can remain a simple winner-takes-all (WTA) choice. The absence of an iteratively updated selection process significantly reduces the computational payload of this step.

The matching cost function utilized is the absolute differences (AD). AD is inherently the simplest metric of all, involving only summations and finding absolute values. Disparity space image (DSI) is a 3D matrix containing the

computed matching costs for every pixel and for all its potential disparity values. The DSI values for constant disparity value are aggregated inside fix-sized square windows. The dimensions of the chosen aggregation window play an important role in the quality of the final result. Generally, small dimensions preserve details but suffer from noise. On the contrast, large dimensions may not preserve fine details but significantly suppress the noise. Noise suppression is very important for stereo algorithms that are intended to be applied to outdoors scenes. Outdoors images, which is often the case for autonomous navigation tasks, usually suffer from noise induced by a variety of reasons, e.g. lighting differences and reflections. The aggregation windows dimensions used in the proposed algorithm are 13x13 pixels. This choice is a compromise between real-time execution speed and noise cancellation. The AD aggregation step of the proposed algorithm is a weighted summation. Each pixel is assigned a weight depending on its Euclidean distance from the central pixel. A 2D Gaussian function determines the weights value for each pixel. The center of the function coincides with the central pixel. The standard deviation is equal to the one third of the distance from the central pixel to the nearest window-border. The applied weighting function can be calculated once and then be applied to all the aggregation windows without any further change. Thus, the computational load of this procedure is kept within reasonable limits. The DSI values after the aggregation are furthered refined by applying 3D cellular automata (CA). Two CA transition rules are applied to the DSI. The values of parameters used by them were determined after extensive testing to perform best. The first rule attempts to resolve disparity ambiguities. It checks for excessive consistency of results along the disparity axis and, if necessary, corrects on the perpendicular plane. The second rule is used in order to smoothen the results and at the same time to preserve the details on constant-disparity planes. The two rules are applied once. Their outcome comprises the enhanced DSI that will be used in order the optimum disparity map to be chosen by a simple, non-iterative WTA step.

Reliability-Based Fusion Algorithm. The output of the utilized custom stereo correspondence algorithm for each image pair is not only a disparity map but a certainty map as well. That is, for every pixel of the disparity map (depth metric for the corresponding pixel in the reference image) a certainty measure is calculated indicating the likelihood of the pixel's selected disparity value to be the right one. The certainty measure is calculated for each pixel (x, y) as can be seen in (1).

$$cert(x,y) = \left| SAD(x,y,D) - \frac{\sum_{z=0}^{d-1} SAD(x,y,z)}{d} \right| \tag{1}$$

According to this, the certainty *cert* for a pixel (x, y) that the computed disparity value D is actually right is equal to the absolute value of the difference between the minimum matching cost value $SAD(x, y, D)$ and the average matching cost

value for that pixel when considering all the d candidate disparity levels for that pixel. What the aforementioned measure evaluates is the amount of differentiation of the selected disparity value with regard to the rest candidate ones. The more the disparity value is differentiated, the most possible it is that the selected minimum is actually a real one and not owed to noise or other effects. The certainty calculation for all the pixels leads to a greyscale certainty map as shown in the third row of Fig. 3 and Fig. 5. In these images the lighter value of a pixel indicates greater certainty about its computed disparity value.

For each pixel, the disparity value having the largest certainty is chosen, among the three candidates. This technique, gets the best parts of each one of the three preliminary disparity maps and fuses them in one final disparity map. The resulting final depth map is obviously more reliable than any of its producing ones.

3 Experimental Results and Applications

The proposed quad-camera algorithm has been applied to two tetrads of images in order to evaluate its performance.

The most common image set for evaluating stereo correspondence algorithms is the Tsukuba data set. While typical stereo algorithms require two horizontally displaced pictures of this set, the proposed method requires four images. The Tsukuba data set consists of multiple images of the scene captured by a camera grid with multiple, equally spaced horizontal and vertical steps. The choice of the images captured by four cameras belonging to a square perfectly satisfies the demands imposed by the proposed algorithm. Figure 3 depicts the reference, i.e. up-left image, in the first column and the three target images, i.e. up-right, down-left and down-right in the second column. The third and the fourth columns show the certainty and disparity maps calculated for the image pairs consisting of the single reference image and the corresponding target ones. The fifth column of the figure shows the fused final disparity map on the top and the ground truth disparity map on the bottom.

Figure 4 shows the experimental results of the proposed quad-camera algorithm (left), the computationally equivalent simple stereo algorithm (middle) and the utilized single stereo algorithm applied on the horizontal stereo pair (right). The first row shows the calculated disparity maps. The second row shows the maps of pixels with absolute computed disparity error bigger than 1 shown in black. Finally, the third row presents maps of signed disparity error where the middle (50%) gray tone equals to zero error. It is obvious that the simple stereo algorithm, shown in the rightmost column suffers from noise. The usual confrontation of this issue is to enlarge the utilized 13x13 pixel aggregation window during the respective stage. However, window enlargement generally leads to loss of detail and coarse results, as shown in the middle column. This version of the algorithm utilizes a 23x23 pixel aggregation window, which results in triple computational load. Obviously, both of these treatments lack the results' quality of the proposed method. The final result of the proposed algorithm requires

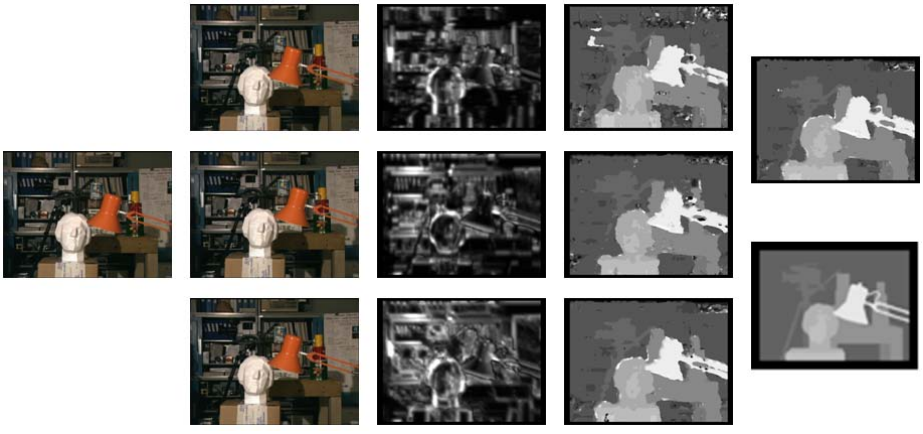

Fig. 3. Algorithm's Steps and Results for the Tsukuba Data Set. (Column 1) the Reference Image (Up-Left), (Column 2) the Three Target Images (Up-Right, Down-Left, Down-Right), (Column 3) the Certainty Maps for the Horizontal, Vertical and Diagonal Pair, (Column 4) the Computed Disparity Map for Each Stereo Pair, (Column 5) the Fused (Top) and the Ground Truth (Bottom) Disparity Maps.

roughly the same computational power as the algorithm in the middle column. The outcome is that the proposed quad-camera algorithm achieves better results than its computationally equivalent simple two-camera stereo counterpart and the simple initial stereo algorithm.

The percentage of pixels whose absolute disparity error is greater than 1 in the non-occluded, all, and near discontinuities and occluded regions are presented in Table 1. The presented percentages refer to the three initially computed stereo pairs (namely the horizontal, vertical and diagonal pair), the final fused result of the proposed system and, finally, the computationally equivalent two-camera utilized stereo correspondence algorithm.

As shown in Table 1 there are cases where the results of the fusion process are marginally worse than those of an initial step. However, the image pair direction that provides the optimum results and should be considered as the most reliable and useful can not be initially anticipated. Moreover, the optimum direction is arbitrary and, therefore, there is little chance to coincide with any of the available three in the proposed system throughout the whole scene. However, the goal of the fusion system is to identify the best disparity value for every pixel. Thus, the results will be roughly as, or occasionally even more, accurate as the best initial results. On the other hand, the final disparity map is, in any case, far more reliable than the initial ones, since it has gone through a validation procedure, guaranteed by (1).

The proposed algorithm has been also applied to a virtual scenery. A virtual quad-camera system was inserted to the virtual room shown in the two first columns of Fig. 5 and the demanded tetrad of images was captured. The room

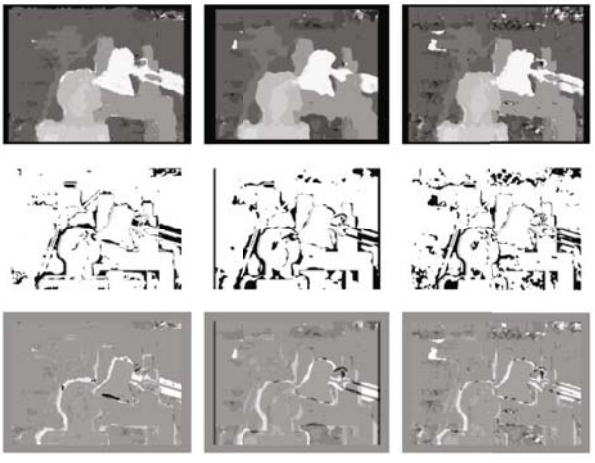

Fig. 4. Results of the Proposed Fusion System (Left), the of the Computationally Equivalent Simple Stereo Algorithm (Middle) and the Preliminary Simple Stereo Algorithm Applied on the Horizontal Image Pair (Right). From Top to Bottom: the Computed Disparity Maps, Pixels with Absolute Computed Disparity Error Bigger than 1 and Maps of Signed Disparity Error.

Table 1. Percentage of Pixels whose Absolute Disparity Error is Greater than 1 in Various Regions for the Tsukuba Pairs

Pair	**Non-occluded (%)**	**All (%)**	**Discontinuities (%)**
Horizontal	16.2	18.1	29.9
Vertical	12.5	13.8	35.1
Diagonal	10.7	12.4	32.3
Proposed	**10.8**	**12.6**	**31.5**
Equivalent	15.8	17.6	33.9

scene was chosen as it is a complex and demanding one, having both regions with fine details and low-textured ones. Moreover, the repetitive pattern of the books, in the background, is a challenging element for the stereo correspondence algorithms. Figure 5 depicts the reference i.e. up-left image in the first column and the three target images i.e. up-right, down-left and down-right in the second column. The third and the fourth columns show the certainty and disparity maps calculated for the image pairs consisting of the single reference image and the corresponding target ones. Finally, the fifth column of the figure shows the fused final disparity map.

The availability of reliable depth maps is the cornerstone of many computer vision as well as robotic applications. Figure 6(a) shows a screenshot of the 3D reconstructed Tsukuba scene. The depth map of Fig. 3 obtained using the

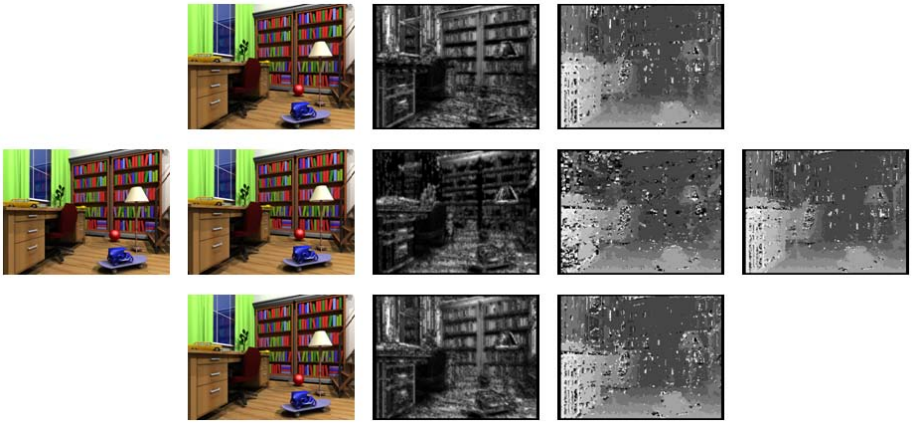

Fig. 5. Algorithm's Steps and Results for a Synthetic Room Scene. (Column 1) The Reference Image (Up-Left), (Column 2) the Three Target Images (Up-Right, Down-Left, Down-Right), (Column 3) the Certainty Maps for the Horizontal, Vertical and Diagonal Pair, (Column 4) the Computed Disparity Map for Each Stereo Pair, (Column 5) the Final Fused Depth Map.

(a) (b)

Fig. 6. Application Results Obtained Using the Caluclated Depth Maps. (a) View of the Reconstructed Tsukuba Scene and (b) Obstacle Detection in the Virtual Room Scene.

proposed method was utilized in order to add the third dimension's information to the reference image. Thus, a 3D model of the scene was reconstructed and a computer user can virtually navigate around the scene. On the other hand, Fig. 6(b) shows an obstacle detection application based on the availability of reliable depth map. Stereo vision can be used by autonomous robotic platforms in order to reliably detect obstacles within their movement range and move accordingly. The previously obtained depth map of Fig. 5 was used for the calculation of the v-disparity image. Using the Hough transformation the floor plane

was calculated and the obstacles were detected. This result is useful for any path-planning algorithm.

4 Conclusions and Discussion

In this work a depth computing system has been proposed aimed for autonomous robotics applications. The system utilizes a square formation of four identical cameras capturing the same scene. Selecting one of the images of each tetrad as reference, three image pairs result. Each pair is processed by a simple and rapid, custom stereo correspondence algorithm which results in an initial disparity map, as well as a certainty map. A fusion process evaluates the three initial disparity maps according to their certainty and produces the final combined disparity map.

Autonomous robotic applications demand reliable depth estimations obtained in real-time frame rates, having at the same time limited computational resources. The proposed system substitutes the computational complexity with a special sensor configuration. However, the demanded configuration can be easily and cost efficiently be achieved. The presented results exhibit a fair compromise between the objectives of low computational complexity and result's reliability.

The accuracy of local algorithms in various regions of a scene is strongly correlated to the orientation of the depicted objects in that particular region towards the orientation of the correspondence search procedure. That is, the depth discontinuities are more discriminable when they are oriented vertically to the correspondence search direction. This conclusion is based on the inherent way local algorithms operate and can be confirmed by the preliminary disparity maps, presented in the fourth row of Fig.3 and Fig. 5. The proposed system has the advantage of being able to adapt to various objects' orientations. The result is that the final fused disparity map is at least as accurate as the most accurate of the initial disparity maps and at the same time much more reliable than any of them. Moreover, the structure of the proposed software architecture is ideal for execution on the nowadays widely available quad-core processors. Each one of the identical but separate stereo correspondence searches can be assigned to a core, while the fourth core will supervise the whole procedure.

Acknowledgments. This work was supported by the E.C. funded research project for vision and chemiresistor equipped web-connected finding robots, "View-Finder", FP6-IST-2005-045541.

References

1. Hariyama, M., Takeuchi, T., Kameyama, M.: Reliable stereo matching for highly-safe intelligent vehicles and its vlsi implementation. In: IEEE Intelligent Vehicles Symposium, pp. 128–133 (2000)
2. Murray, D., Little, J.J.: Using real-time stereo vision for mobile robot navigation. Autonomous Robots 8(2), 161–171 (2000)

3. Sim, R., Little, J.J.: Autonomous vision-based robotic exploration and mapping using hybrid maps and particle filters. Image and Vision Computing 27(1-2), 167–177 (2009)
4. Schreer, O.: Stereo vision-based navigation in unknown indoor environment. In: Burkhardt, H.-J., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1406, pp. 203–217. Springer, Heidelberg (1998)
5. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47(1-3), 7–42 (2002)
6. Torra, P.H.S., Criminisi, A.: Dense stereo using pivoted dynamic programming. Image and Vision Computing 22(10), 795–806 (2004)
7. Muhlmann, K., Maier, D., Hesser, J., Manner, R.: Calculating dense disparity maps from color stereo images, an efficient implementation. International Journal of Computer Vision 47(1-3), 79–88 (2002)
8. Di Stefano, L., Marchionni, M., Mattoccia, S.: A fast area-based stereo matching algorithm. Image and Vision Computing 22(12), 983–1005 (2004)
9. Yoon, S., Park, S.K., Kang, S., Kwak, Y.K.: Fast correlation-based stereo matching with the reduction of systematic errors. Pattern Recognition Letters 26(14), 2221–2231 (2005)
10. Nalpantidis, L., Sirakoulis, G.C., Gasteratos, A.: Review of stereo vision algorithms: from software to hardware. International Journal of Optomechatronics 2(4), 435–462 (2008)
11. Ruigang, Y., Welch, G., Bishop, G.: Real-time consensus-based scene reconstruction using commodity graphics hardware. In: 10th Pacific Conference on Computer Graphics and Applications, pp. 225–234 (2002)
12. Schirmacher, H., Li, M., Seidel, H.P.: On-the-fly processing of generalized lumigraphs. In: EUROGRAPHICS, pp. 165–173 (2001)
13. Yang, J.C., Everett, M., Buehler, C., Mcmillan, L.: A real-time distributed light field camera. In: Eurographics Workshop on Rendering, pp. 77–86 (2002)
14. Wilburn, B., Smulski, M., Lee, K., Horowitz, M.A.: The light field video camera. In: Media Processors, pp. 29–36 (2002)
15. Zitnick, C.L., Kanade, T.: A cooperative algorithm for stereo matching and occlusion detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7), 675–684 (2000)
16. Mingxiang, L., Yunde, J.: Trinocular cooperative stereo vision and occlusion detection. In: IEEE International Conference on Robotics and Biomimetics, pp. 1129–1133 (December 2006)
17. Nalpantidis, L., Sirakoulis, G.C., Gasteratos, A.: A dense stereo correspondence algorithm for hardware implementation with enhanced disparity selection. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds.) SETN 2008. LNCS (LNAI), vol. 5138, pp. 365–370. Springer, Heidelberg (2008)