Robust 3D Vision for Robots Using Dynamic Programming

Lazaros Nalpantidis[1], John Kalomiros[2], and Antonios Gasteratos[1]

[1]Laboratory of Robotics and Automation, Department of Production and Management Engineering
Democritus University of Thrace, Vas. Sofias 12, GR-67100, Xanthi, Greece
Email: {lanalpa,agaster}@pme.duth.gr.
[2]Department of Informatics and Communications, School of Technological Applications
Technological Educational Institute of Serres, Terma Magnisias, GR-62124, Serres, Greece
Email: ikalom@teiser.gr

***Abstract*—In this paper a new stereo vision method is presented that combines the use of a lightness-invariant pixel dissimilarity measure within a dynamic programming depth estimation framework. This method uses concepts such as the proper projection of the HSL colorspace for lightness tolerance, as well as the Gestalt-based adaptive support weight aggregation and a dynamic programming optimization scheme. The robust behavior of this method is suitable for the working environments of outdoor robots, where non ideal lighting conditions often occur. Such problematic conditions heavily affect the efficiency of robot vision algorithms in exploration, military and security applications. The proposed algorithm is presented and applied to standard image sets.**

***Index Terms*—stereo vision, robot vision, dynamic programming, lightness-invariant.**

I. Introduction

Autonomous robots need to know about the structure of their 3D environment as it plays a decisive role in their behavior and planning. Vision is an intuitive way to gather information about the world. Furthermore, stereo vision is able to extract the depth of a scene out of two images. However, stereo correspondence is considered as a demanding and computationally intensive procedure. The recent advances in stereo vision algorithms [1] have made such systems suitable for robots. On the other hand, robotics poses new problems to the stereo vision algorithms. Situations of non-uniform illumination often occur in real working environments, as shown in Fig. 1. As a result, apart from high refresh rates and precise results for ideal image pairs, stereo algorithms should be also able to cope with difficult illumination conditions [2].

There are two large families of stereo algorithms, i.e. local and global ones [1], [3]. Local stereo algorithms can provide high frame rates but their accuracy is low. On the other hand, global algorithms suffer from low frame rates but their results are generally very accurate. Dynamic Programming (DP) stands somewhere between those two broad classes providing good accuracy of results in acceptable frame rates. Moreover, recent advances in DP-based stereo algorithms seem to be able to significantly improve both of these two aspects. Hardware implementations of DP have been reported [4], [5] that provide high execution speed. Additionally, the incorporation of adaptive support weight aggregation (ASW) schemes has been shown to further improve the accuracy and detail of the produced depth maps [6].

Fig. 1. Left and right images of pairs suffering from non-uniform illumination

II. Algorithm Description

This work presents a new stereo algorithm that uses a lightness-invariant pixel dissimilarity measure, a ASW-based aggregation scheme, and a DP-based optimization step. The overall structure of the presented algorithm is shown in Fig. 2.

Fig. 2. Block diagram of the presented stereo correspondence algorithm

A. Lightness Compensating Dissimilarity Measure

The used dissimilarity measure is defined and calculated within the HSL colorspace. This colorspace is represented as a double cone. The H channel is for hue and expresses the human impression about which color is actually depicted. Each color is represented by an angular value ranging between 0 and 360 degrees (0 being red, 120 green and 240 blue). The S channel is for saturation and determines how vivid or gray the particular color is shown. Its value ranges from 0 for gray to 1 for fully saturated (pure) colors. Finally, the L channel of the HSL colorspace is for the Luminosity and it determines the intensity of a specific color. It ranges from 0 for completely dark colors (black) to 1 for fully illuminated colors (white).

As a result, the HSL colorspace separates lightness from the other pure chromatic characteristics. This fact implies that a given color will theoretically result in the same values of hue and saturation regardless the environment's illumination conditions. Ignoring the Luminosity channel will inevitably lead to the loss of some amount of information and, as a result, to slightly inferior results for ideal lighting, but will provide robustness against real, non-ideal and non-uniform lighting conditions [7]. The omission of the vertical (L) axis from the colorspace representation leads to a 2D circular disk, defined only by H and S.

In this reduced colorspace, each color $\mathbf{P}_i$ can be represented as a planar vector with its initial point being the disc's center. As a consequence, it can be described as a polar vector or equivalently as a complex number with modulus equal to S_i and argument equal to H_i. That is, a color in the new luminosity-ignoring colorspace representation can be described as:

$$\mathbf{P}_i = S_i e^{iH_i} \tag{1}$$

Based on this color description a luminosity-compensated dissimilarity measure (LCDM) has been proposed in [7]. According to this, the variance of two colors $\mathbf{P}_1$ and $\mathbf{P}_2$ can be found in the reduced HS colorspace as the difference of the two complex numbers:

$$\begin{aligned} LCDM_{P_1,P_2} &= |\mathbf{P}_1 - \mathbf{P}_2| \\ &= \left|S_1 e^{iH_1} - S_2 e^{iH_2}\right| \\ &= \sqrt{{S_1}^2 + {S_2}^2 - 2S_1S_2\cos(H_1 - H_2)} \end{aligned} \tag{2}$$

Equation 2 is the mathematical formulation of the LCDM dissimilarity measure, which takes into consideration any pure chromatic information, but not the luminosity. In contrast to other popular dissimilarity measures, such as absolute differences (AD) or squared differences (SD), LCDM can provide robust behavior against viewpoint-dependent chromatic differentiations. Consequently, LCDM was chosen to be used in this algorithm.

B. Gestalt-based Aggregation

Aggregation is used as a mean to suppress the existence of noise during the subsequent disparity value selection. Commonly, the dissimilarity values of all the pixels $B(x', y')$ lying inside a $w \times w$ support region around a central pixel $A(x, y)$ for a given disparity value d are aggregated as the updated value of pixel A for the considered disparity value. While fix and adaptive sized support windows have been commonly used for constant-weight pixel aggregation, ASW [8] has proposed the use of fix sized windows and adaptively weighted pixel values. The weight assignment can be performed according to the Gestalt laws of perceptual organization. The proposed algorithm follows a mathematical formulation similar to the one found in [7] for the Gestalt laws of proximity and similarity. According to that, the two Gestalt laws can be expressed as:

- Proximity (or equivalently distance): The closer two pixels are the more correlated to each other they are.

$$proximity_{A,B} = 1 - \frac{\sqrt{(x - x')^2 + (y - y')^2}}{w\sqrt{2}} \tag{3}$$

- Color similarity (or equivalently color dissimilarity): The more similar the colors of two pixels are the more correlated they are.
  The color similarity of the two pixels can be estimated using the LCDM of their colors. Thus, the similarity between the pixels A and B is calculated as:

$$similarity_{A,B} = 1 - \frac{LCDM_{(x,y),(x',y')}}{2} \tag{4}$$

Finally, an ASW aggregation scheme requires a function that combines the quantified Gestalt laws in order to provide a single weighting factor. The two Gestalt-based correlation factors are combined into one by multiplication providing a general correlation weight between the pixels A and B:

$$w_{A,B} = proximity_{A,B} \cdot similarity_{A,B} \tag{5}$$

This combined weight of Eq. 5 is calculated for the support regions of the examined pixels both in the left and the right input images, obtaining $w_{A,B_{left}}$ and $w_{A,B_{right}}$, respectively. Consequently, the aggregation of the LCDM, taking into consideration the weighting factor for each pixel is:

$$ASW_A = \frac{\sum w_{A,B_{left}} \cdot w_{A,B_{right}} \cdot LCDM_{A,B}}{\sum w_{A,B_{left}} \cdot w_{A,B_{right}}} \tag{6}$$

where the pixel B belongs to the $w \times w$ neighborhood of the central pixel A.

C. Dynamic Programming (DP) Optimization

Dynamic programming has been used widely as a semi-global optimization method for the estimation of disparity d along image scanlines [9], [10]. The general idea behind a DP stereo algorithm is to treat the correspondence problem as an energy minimization problem. The energy function E represents the total cost of a sequence M of matching pixels in a scanline and consists of a data and smoothness term, according to the equation:

$$E(M) = E_{data}(M) + E_{smooth}(M) \quad (7)$$

The data term, in the above equation is a dissimilarity measure between pixels in the left and right image of the stereo pair. Intensity differences have been used widely in the literature, while in this paper we introduce the luminosity-compensated dissimilarity measure (LCDM) presented in Eq. 2 and aggregated using the adaptive weight support scheme, as in Eq. 6. The smoothness term can be formulated to handle depth discontinuities and occlusions. The total cost function of a matching sequence, used in this paper, originates from Birchfield and Tomasi's empirical formulation [11], which is found to facilitate the precise localization of depth discontinuities:

$$E(M) = \sum_{i=1}^{N_m} ASW(x_i, y_i) + N_{occ}k_{occ} - N_m k_r \quad (8)$$

where k_{occ} is a constant occlusion penalty, k_r is a constant match reward, ASW is the aggregated luminocity-compensated dissimilarity measure between two pixels x_i, y_i in the left and right scanline, and N_{occ}, N_m are the number of occlusions and matches, respectively, in sequence M [11].

In our approach of DP a $2D$ cost plane is built for all possible disparities, for each pair of horizontal image lines. Fig. 3 shows the cost plane for a maximum disparity range of five pixels. A cost is attributed to every cell in the grid and the algorithm searches for the path of minimum total cost. Gray cells do not belong to the search grid because they are beyond the limits of maximum disparity. Marked cells represent the matching sequence in the presented example. Columns or rows without marked pixels correspond to an occluded pixel.

A different representation of the cost plane can result by shifting up each column of Fig. 3 by an amount equal to the column index, as shown in Fig. 4. In the grid of Fig. 4 the vertical axis represents the disparity $d = x - y$. Taking into account the ordering and uniqueness constraints and assuming that occlusions cannot occur simultaneously in the left and right scanlines, Birchfield and Tomasi suggest that for any cell in the grid of Fig. 4 the possible preceding matches are the cells shown in Fig. 5. For each cell in the grid of Fig. 4 we record the cost $\phi(d, y)$, which represents the cost of the best match sequence up to the present point. The ϕ array is traversed from left to right and from top to bottom and the cost of the best path to each cell is computed as in [11]:

$$\phi[d, y] = ASW(y + d, y) - k_r + \min \left\{ \begin{array}{l} \phi[d, y-1], \\ \phi[d-1, y-1] + k_{occ}, ..., \\ \phi[0, y-1] + k_{occ}, \\ \phi[d+1, y-2] + k_{occ}, ..., \\ \phi[d_{max}, y + d - d_{max} - 1] + k_{occ} \end{array} \right\} \quad (9)$$

In Eq. 9 the minimum is taken among all possible preceding matches, shown as cells in Fig. 5. The first cost corresponds to a match without change in disparity, while all other costs precede left and right occlusions. The value d_{max} represents the upper limit of disparity.

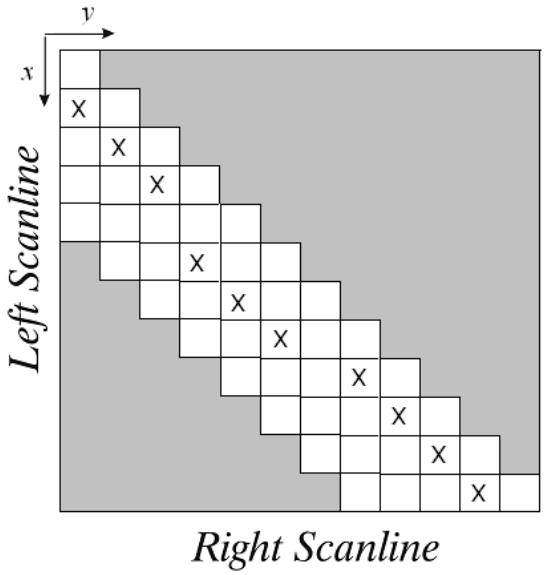


Fig. 3. Cost plane shown as a search grid for a left and right scanline, with a maximum disparity range of five pixels. Marked cells form a match sequence and unmarked columns or rows correspond to occlusions

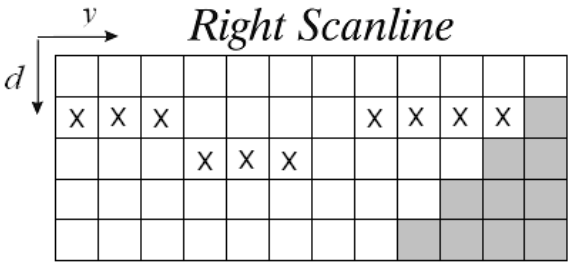


Fig. 4. Shifted cost plane, where the vertical axis corresponds to disparity $d = x - y$. Marked cells represent the same match sequence, as in Fig. 3

A second matrix π is also filled for each grid cell, where each cell $\pi[d, y]$ contains the coordinates $[d_p, y_p]$ of the immediately preceding match in the match sequence.

After building the cost-plane, the optimal path is found by backtracking. Starting from the lowest cost cell corresponding to a limiting pixel in the right scanline we trace the optimal path by following backwards the cells of the array π. All the unmatched pixels in an occlusion are assigned the disparity of the nearest match.

Let us note that the above formulation leads to a multi-state dynamic programming approach, as opposed to the usual three-state approach. In the later, for each cell in the search grid there are only three candidate matches preceding the current cell [9].

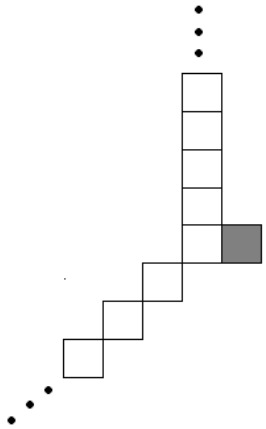


Fig. 5. For each cell in the grid of Fig. 4, displayed here as a gray square, there is a set of immediately preceding possible matches, shown here as white grid cells

By applying the proposed luminocity-compensated matching cost in the cost-computation stage described by Eq. 9, we explore the use of an intensity-invariant measure in a dynamic programming framework. Additionally, the proposed adaptive weighted cost aggregation step selectively acquires interscanline support from adjacent scanlines. As a result, it is expected to improve the overall dense disparity map by reducing the horizontal streaking artifacts usually present in disparity maps produced by dynamic programming techniques. Experimental results obtained from public stereo datasets are presented in the next section.

III. Experimental Results

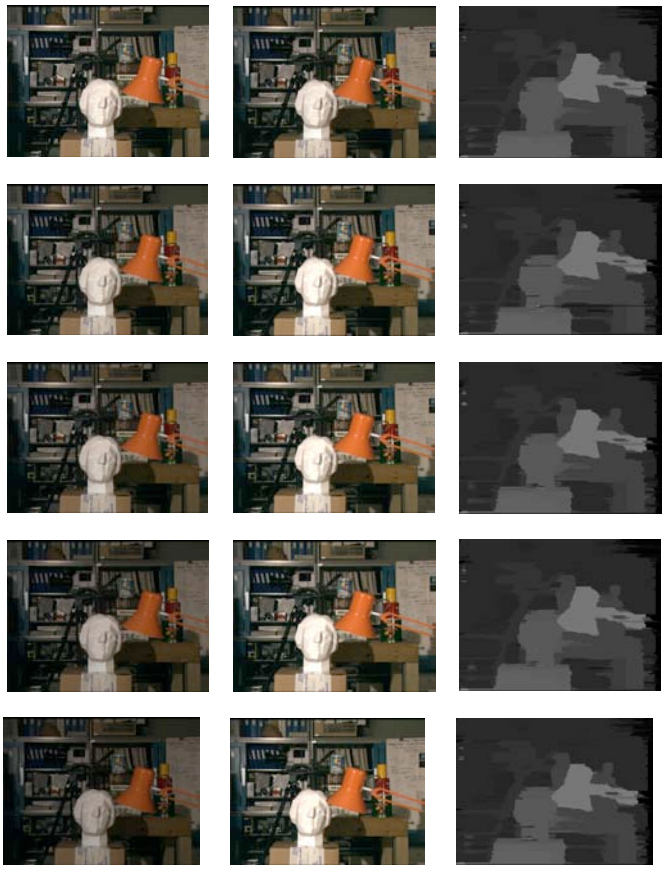
The proposed algorithm was tested on a series of stereo image pairs in order to assess its performance. The datasets used for the test were the popular and widely used by the stereo vision community Tsukuba and Cones image pairs. These sets of images are known for the combination of regions with different characteristics and are challenging for stereo vision algorithms.

The parameters of the proposed algorithm were given constant values throughout the experimental validation tests. An aggregation window of 9×9 pixels was used. The occlusion penalization parameter has been set to $k_{occ} = 5$ and the constant match reward parameter was $k_r = 25$. The results of the proposed algorithm when applied to the ideally lightened Tsukuba and Cones image pairs are given in Fig. 6.

The proposed algorithm has been also applied to a series of stereo pairs suffering from differences of the lightness between the two input images. The differently lightened image pairs were derived based on the original Tsukuba stereo pair and by altering the values of each one of the RGB channels by a fixed percent each time. As a result, a series of Tsukuba pairs have been obtained, shown in Fig. 7(a) and 7(b), with a fixed lightness differentiation between the two images of each pair. The proposed algorithm was tested on these image pairs and the resulting disparity maps are given in Fig. 7(c).

(a) Left image (b) Right image (c) Disparity map

Fig. 6. Final disparity maps produced by the proposed algorithm for the Tsukuba (first row) and the Cones (second row) datasets under perfect illumination conditions



(a) Left image (b) Right image (c) Disparity map

Fig. 7. Final disparity maps produced by the proposed algorithm for the Tsukuba dataset under 0% (first row), 20% (second row), 30% (third row) , 40% (fourth row) and 50% (fifth row) illumination differentiation

Table I shows the normalized mean square error (NMSE) of the calculated disparity map for each of the pairs of Fig. 7 with respect to the ground-truth disparity maps [3], [12].

Using least squares fitting, it is obtained that the trendline's slope for the values given in Table I is $-2 \cdot 10^{-5}$. This proves that the output of the algorithm is largely independent of the illumination difference between the two input images. As a

TABLE I
NMSE FOR VARIOUS ILLUMINATION DIFFERENCES OF THE INPUT IMAGES

Illumination Difference	NMSE
0%	0.0660
20%	0.0665
30%	0.0679
40%	0.0652
50%	0.0649

result, the algorithm presents a robust behavior having low NMSE constantly over a wide range of lightness differentiations between the two input images.

IV. CONCLUSIONS

This work has presented a new robust stereo vision algorithm suitable for use in a robot's real working environment. The use of the lightness-invariable dissimilarity measure LCDM, a sophisticated Gestalt-based ASW aggregation procedure, and a DP-based optimization step is proposed. The algorithm can tolerate non-ideally illuminated input images and produce quality and reliable results. In order to evaluate the proposed algorithm the public Middlebury dataset [12] were used. The results are very encouraging and show that good disparity maps can be obtained even under strongly biased lightness differences between the images of the stereo pair. In addition, the proposed adaptive cost aggregation scheme strongly eliminates horizontal streaks that are due to inconsistencies between successive scanlines.

In future work the proposed DP framework will be studied with respect to its hardware implementation as a System-on-a-chip, with the purpose of high quality depth maps in very high frame-rates. First results show that the system can successfully be implemented using medium hardware resources in a field programmable gate array (FPGA) chip.

REFERENCES

[1] L. Nalpantidis, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: from software to hardware," *International Journal of Optomechatronics*, vol. 2, no. 4, pp. 435–462, 2008.

[2] G. Klancar, M. Kristan, and R. Karba, "Wide-angle camera distortions and non-uniform illumination in mobile robot tracking," *Journal of Robotics and Autonomous Systems*, vol. 46, pp. 125–133, 2004.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[4] J. A. Kalomiros and J. Lygouras, "Hardware implementation of a stereo co-processor in a medium-scale field programmable gate array," *IET Computers and Digital Techniques*, vol. 2, no. 5, pp. 336 –346, 2008.

[5] J. Kalomiros and J. Lygouras, "Comparative study of local sad and dynamic programming for stereo processing using dedicated hardware," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–18, 2009.

[6] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister, "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," in *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, 2006, pp. 798–805.

[7] L. Nalpantidis and A. Gasteratos, "Stereo vision for robotic applications in the presence of non-ideal lighting conditions," *Image and Vision Computing*, vol. 28, pp. 940–951, 2010.

[8] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, 2006.

[9] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs, "A maximum likelihood stereo algorithm," *Computer Vision and Image Understanding*, vol. 63, pp. 542–567, 1996.

[10] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, pp. 181–200, 1999.

[11] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 269–293, December 1999.

[12] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 195–202.