



Anantrasirichai, N., Canagarajah, C. N., Redmill, D. W., & Bull, D. R. (2006). Dynamic programming for multi-view disparity/depth estimation. In 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006) Toulouse, France. (Vol. 2, pp. 269 - 272). Institute of Electrical and Electronics Engineers (IEEE).
10.1109/ICASSP.2006.1660331

Link to published version (if available):
[10.1109/ICASSP.2006.1660331](https://doi.org/10.1109/ICASSP.2006.1660331)

[Link to publication record in Explore Bristol Research](#)
PDF-document

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/pure/about/ebr-terms.html>

Take down policy

Explore Bristol Research is a digital archive and the intention is that deposited content should not be removed. However, if you believe that this version of the work breaches copyright law please contact open-access@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline of the nature of the complaint

On receipt of your message the Open Access Team will immediately investigate your claim, make an initial judgement of the validity of the claim and, where appropriate, withdraw the item in question from public view.

DYNAMIC PROGRAMMING FOR MULTI-VIEW DISPARITY/DEPTH ESTIMATION

N. Anantrasirichai, C. Nishan Canagarajah, David W. Redmill and David R. Bull

Department of Electrical & Electronic Engineering,
University of Bristol, Bristol BS8 1UB UK

ABSTRACT

A novel algorithm for disparity/depth estimation from multi-view images is presented. A dynamic programming approach with window-based correlation and a novel cost function is proposed. The smoothness of disparity/depth map is embedded in dynamic programming approach, whilst the window-based correlation increases reliability. The enhancement methods are included, i.e. adaptive window size and shiftable window are used to increase reliability in homogenous areas and to increase sharpness at object boundaries. First, the algorithm estimates depth maps along a single camera axis. The algorithm exploits then combines the depth estimates from different axis to derive a suitable depth map for multi-view images. The proposed scheme outperforms existing approaches in parallel and in the non-parallel camera configurations.

1. INTRODUCTION

The disparity/depth estimation algorithm from the sequences of stereo/multi-view images is an important element in 3D vision. Owing to occlusions, imperfect camera calibrations, imperfect light balance and homogenous colour/luminance, the accurate estimate of disparity/depth remains a challenging problem. In this paper, we use dynamic programming to determine the optimal disparity/depth fields, which was first introduced in [1] for stereo images. The dynamic programming approach is better than traditional matching schemes, since it does not contain blocking artefacts or noisy depth maps. Moreover, the dynamic programming is efficient to solve multi-stage problems, which enables disparity estimation and occlusion detection simultaneously.

The choice of a good cost functions for searching the minimum-cost path is a key aspect of the dynamic programming approach. The simplest cost function utilizes the similarity of luminance between the left and the right views. I. Cox has proposed the matching process using individual pixel intensity [2]. Although cohesivity constraints are used to deal with the inter-scanlines disparity discontinuities, the ambiguity from imperfect light balance might affect homogenous areas. Therefore, we investigate the performance of the window-based correlation for dynamic programming and propose some enhancements such as adaptive window size and shiftable windows. The small window size and shiftable windows lead to

sharp object boundaries, whilst the large window size achieves better matching in homogeneous areas [3].

The dynamic programming scheme can be improved by supplementing the cost function to identify the occlusion. The accuracy of the occlusion detection has been enhanced by Bayesian method which requires a probability of occlusion [4,5]. N. Grammalidis and M. G. Strinzi have proposed the disparity estimation and the occlusion detection algorithm for multi-view system by using dynamic programming, but cost defined to identify occlusion is fixed [6]. In this paper, we propose a simple but effective cost function by exploiting confidence information from other cameras. The erroneous prediction by using one view reference is avoided by exploiting the other reference views. This produces better disparity estimation by using multiple reference views.

The proposed scheme is tested with the parallel and non-parallel camera configurations. Each configuration is considered with two types of camera geometries: a linear case (one dimension) and a planar case (two dimensions). In the later case, a joint horizontal and vertical scanning is proposed to deal with the multiple global minima. The rest of paper is organized as follows: Section 2 describes the proposed cost functions. Then, Section 3 explains the window-based correlation with adaptive window size and shiftable window. The performance of the proposed scheme is tested with two camera configurations and results are presented in Section 4 with conclusions in Section 5.

2. PROPOSED COST FUNCTIONS FOR DYNAMIC PROGRAMMING

We propose to search the matching pixels along each scanline through dynamic programming and the disparity/depth of each pixel comes from the optimum path. Considering one particular pixel, there are three possible disparity values, which are equal to, more or less than that of the consecutive pixel. The first case usually occurs in the non-occlusion areas, whilst the last two cases possibly occur in the half-occlusion areas. Hence, three costs are defined to each node (i, j) in dynamic programming; C_1 and C_2 are the occlusion costs of the pixel, which are invisible in left view and right view respectively, and C_3 is the cost of the pixel in non-occlusion areas. These costs are expressed as follows:

$$C_1 = \left(\frac{1}{\lambda} - E_l \right) + E_r, \quad C_2 = \left(\frac{1}{\lambda} - E_r \right) + E_l, \quad C_3 = E_l + E_r \quad (1)$$

$$E_l = \frac{e_{ij}^2(d)}{1 + \lambda \cdot e_{ij}^2(d)}, \quad E_r = \frac{e_{ij}^2(\frac{B^{R_1}}{B^{L_1}}d)}{1 + \lambda \cdot e_{ij}^2(\frac{B^{R_1}}{B^{L_1}}d)} \quad (2)$$

where $e_{ij}^2(k) = \sum_{(i,j) \in w_{ij}} a_{ij} \cdot (p_{ij}^c - p_{(i)(j+k)}^r)^2$, a_{ij} is a weighing

coefficient that is inversely proportional to the distance from the pixel (i,j) , and p_{ij} is the intensity of pixel (i,j) . The B^{L_1} and B^{R_1} are the baseline between the current camera and the closest left and the closest right cameras respectively, whilst λ is the occlusion parameter.

The occlusion parameter directly affects the costs in dynamic programming. The appropriate value depends upon the details of images, i.e. the multi-view image sequences composed of a wide range of depths require a large λ for operating and vice versa. The areas where the depth is changing usually contain the high different luminance from the reference view at the same position, so an initial λ is determined as follows. For each pixel (i,j) , the energy of different luminance between the current and the reference view, $e_{ij}^2(0)$, is computed.

Then, the value $1/\lambda$, which is the intersection of function E and $(1/\lambda - E)$, is defined to be an energy threshold that half the number of pixels of an image have the energy $e_{ij}^2(0)$ less than $1/\lambda$. We suggest that the experiments would run with the λ values near the initial one to find the best result.

The accumulated cost of each node (i, j) is marked as in Eq.3 and the selected path of each node is in Eq.4.

$$\text{cost}(i, j) = \min(\text{cost}(i-1, j) + C_1, \text{cost}(i, j-1) + C_2, \dots, \text{cost}(i-1, j-1) + C_3) \quad (3)$$

$$\text{path}(i, j) = \arg \min_d \text{cost}(i, j) \quad (4)$$

Finally, the least accumulated cost of the last pixel in a scanline is selected to identify the optimum path. After tracking back along the optimum path, the estimated disparities are generated and the occlusion areas are simultaneously marked at the pixels where the occlusion cost dominates.

The proposed algorithm achieves the disparity estimation and occlusion detection by exploiting the error of the intensity/colour matching. The large $e^2(d)$ value implies occlusion that the current pixel could not match any pixel in the reference view. This situation always occurs near the object boundaries that the disparity of this pixel is different from those of neighbouring pixels. Therefore, it can be marked that this pixel is located in the background of the current view and it is occluded by foreground in the reference view. That is, if C_1 or C_2 dominates, this pixel is in the background and the disparity is decreasing or increasing respectively, but if C_3 dominates, this pixel could be either in the background or the foreground with the same disparity as the previous pixel. For example, a large E_l and a small E_r cause the value of C_1 less than C_2 and C_3 , therefore C_1 dominates, i.e. this current pixel does not correspond to any pixel of the left view reference but evidently match one pixel in the right view reference. Hence, this pixel is in the background on the right of the foreground object.

Noticeably, the incorrect path might appear at the occlusion

areas composed of similar details of the neighbouring non-occlusion areas. However, if more cameras are available, this problem could be eliminated by using other reference views. The following subsection 2.1 and 2.2 show the proposed schemes for more than three cameras that are available in a linear case and a planar case respectively.

2.1 Multi-view image extension

To extend the three-view disparity estimation to the general multi-view disparity estimation for a linear configuration, the error from all the reference views are compared and the minimum is selected. The error function becomes:

$$E_l = \min \left(\frac{e^2(d)}{1 + \lambda \cdot e^2(d)}, \frac{e^2(\frac{B^{L_2}}{B^{L_1}}d)}{1 + \lambda \cdot e^2(\frac{B^{L_2}}{B^{L_1}}d)}, \dots, \frac{e^2(\frac{B^{L_n}}{B^{L_1}}d)}{1 + \lambda \cdot e^2(\frac{B^{L_n}}{B^{L_1}}d)} \right) \quad (5)$$

$$E_r = \min \left(\frac{e^2(\frac{B^{R_1}}{B^{L_1}}d)}{1 + \lambda \cdot e^2(\frac{B^{R_1}}{B^{L_1}}d)}, \frac{e^2(\frac{B^{R_2}}{B^{L_1}}d)}{1 + \lambda \cdot e^2(\frac{B^{R_2}}{B^{L_1}}d)}, \dots, \frac{e^2(\frac{B^{R_m}}{B^{L_1}}d)}{1 + \lambda \cdot e^2(\frac{B^{R_m}}{B^{L_1}}d)} \right) \quad (6)$$

where B^{L_x} and B^{R_x} are the baselines between the current camera and the x^{th} camera on the left and right sides with total n left view references and m right view references respectively.

Better prediction is achieved by exploiting the information from other view references. However, these error functions are adapted to the cost C_3 only, whilst the cost C_1 and C_2 are still calculated from the closest left and the closest right views. This is because the smaller error might lead the mistaken path in occlusion areas, i.e. C_3 would dominate instead of C_1 or C_2 .

2.2 Combining horizontal and vertical scanning

If the image sequences are available both vertical and horizontal direction, the estimated disparity from scanning in one camera axis could be used to modify the cost for each node of scanning process in another camera axis by pre-marking the occlusion areas. After vertical scanning, for example, the possible occlusion regions of the horizontal scanning are marked from the vertically estimated disparity map, and then the costs C_1 and C_2 of such regions are adjusted with the proportion α as follows:

$$C_1 = \alpha \cdot \left(\left(\frac{1}{\lambda} - E_l \right) + E_r \right), \quad C_2 = \alpha \cdot \left(\left(\frac{1}{\lambda} - E_r \right) + E_l \right) \quad (7)$$

where $0 < \alpha < 1$. As a result, the disparity and occlusion in this horizontal scanning are more reliable than that does not exploit information from the vertical scanning.

Then, the result of the vertical scanning where C_3 dominates is used to replace the result of the horizontal scanning which is marked as occlusions. This could compensate the faults of the one-directional scanning which come from the continuity of the tracked path when the disparity of this area is dropping from the foreground or come from the too narrow gap between the front objects as shown in Fig. 1 at A and B points respectively. In conclusion, to increase the reliability, we also propose to use the result of the first scanning if C_3 s of these areas dominate to replace the result of the second scanning in the occlusion area.

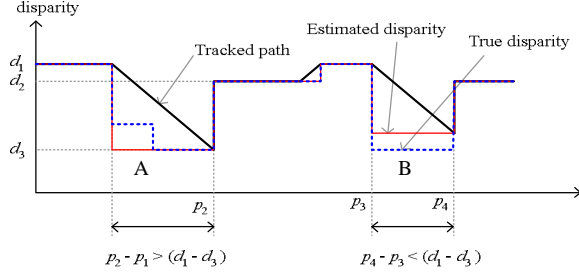


Fig. 1 The disparity estimation with the thick solid line showing the tracked path of the optimum global cost, and the thin solid line showing the estimated disparity got from the tracked path. Some mistakes possibly appear as the difference from the true disparity (dotted line).

3. WINDOW-BASED CORRELATION AND ENHANCEMENT SCHEMES

The window-based correlation is investigated in this paper. The cost of each node is found by exploiting the correlation of the current pixel and its neighbourhoods. The details in a window gain more reliable matching compared to a single pixel. Moreover, the performance of the window-based correlation can be improved by exploiting variable window sizes and shiftable windows method.

The purpose of the adaptive window size scheme is to solve the error matching in the homogenous region. This lies in the basic idea that the higher variance gains the higher perceived reliability. Hence, the window size is grown to cover enough details of the texture, but not too big to make the depth inhomogeneous in a window. In this paper, an adaptive window or block size is chosen based on the following approach. Firstly, each pixel starts with 3x3 window size. Then, if there is no edge inside, the size is increased to be 5x5 pixels. On the contrary, the size of the window that includes the edge details will be decreased to 2x2, if the variance of the window is more than the specific threshold. The results will be shown in the next section.

For the shiftable windows, this algorithm improves the matching areas near the object boundaries that produce the depth discontinuities. The window of the conventional approach is centred at the current pixel, whilst the shiftable window could be located at the appropriate area that minimizes the overlap region. The proper window position might be defined by various criteria. The minimum sum of the absolute difference (SAD) is one of the simple criteria. It is assumed that the homogenous luminance regions contain the homogenous depth. The minimum SAD between the matching windows located at the homogenous depth is less than the minimum SAD between the matching windows located at the area composed of various depths.

4. DISPARITY/DEPTH ESTIMATION FOR MULTI-VIEW IMAGE SYSTEM

The proposed algorithm exploits the property of multi-view points and possible occlusions to indicate the constant or transition stage of disparity/depth. This section shows the results of the proposed scheme described in previous sections.

Firstly, the proposed scheme is applied to the parallel camera configuration, then to non-parallel camera configuration.

4.1. Parallel Camera Configuration

Considering the parallel camera configuration, which is the simplest formation of multi-cameras, the camera axis is not rotated from each other, as well as the world coordinates. If the image point \mathbf{x}_1 is normalized, the space point's coordinates are $\mathbf{X}_w = \mathbf{Z}_w \mathbf{K}^{-1} \mathbf{x}_1 = \mathbf{x}_1 + \mathbf{K} \mathbf{t} / \mathbf{Z}_w$. If the disparity, $d_{1 \rightarrow 2}$, is defined as the vector from a point in image 1 to its corresponding point in image 2, the depth Z_w has a relationship with the disparity $d_{1 \rightarrow 2} = -\mathbf{K} \mathbf{t} / \mathbf{Z}_w$. Additionally, if the translation parameters are reduced to only one direction, the disparity can be rewritten to $d_{1 \rightarrow 2} = -f B_{12} / Z_w$, where f and B_{12} represent the focus length and the baseline between camera 1 and camera 2 respectively. It implies that only the disparity or the shift distance between the corresponding pixels is adequate to represent the 3D information. Therefore, the disparity estimation scheme is proposed for this simple geometry instead of the depth estimation.

The *Head* sequence is employed to simulate the proposed scheme explained in section 3 and 4. The estimated disparity map of the traditional fix window size is illustrated in Fig. 2 (c), whilst the results of the enhanced variable window size scheme is shown in Fig. 2 (d) that is the result of three window sizes as the algorithm described in section 3. Then, the shiftable window approach was tested with fix window size. The estimated disparity map is significantly improved as shown in Fig. 2 (e). Fig. 2 (f) displays the estimated disparity from the vertical scanning by applying both adaptive window size and the shiftable window. Subsequently the predicted occlusion areas for the horizontal scanning, shown in Fig. 2 (g), are marked by the result in Fig. 2 (f). Finally, the disparity field estimated by congregating the horizontal and the vertical scanning as explained in section 3 is illustrated in Fig. 2 (h).

4.2. Non-Parallel Camera Configuration

The disparity of this system is not just the one directional shifting, along neither horizontal nor vertical axis. However, searching the corresponding points among views can be minimized to one direction along the epipolar line.

The epipolar line, \mathbf{E}_v , in view v ($v \in \{l : \text{left view}, r : \text{right view}\}$) that corresponds to the normalized pixel $\mathbf{x}_m = (i, j, 1)^T$ of the middle view m can be calculated by $\mathbf{E}_v = \mathbf{F}_v \mathbf{x}_m$ [7], where \mathbf{F}_v is the fundamental matrix of view v corresponding to the middle view.

Fig. 3 (a) illustrates the original image of *Leo*¹ sequence in the middle view; whilst Fig. 3 (b) and (c) shows the estimated depth map with precision of 0.0076. This 480x270-size sequence is composed of 5 views in the horizontal direction and 3 views in the vertical direction with the non-parallel geometry. From Fig. 3 (b), the result of dynamic programming without any enhance scheme shows the performance of proposed cost

¹ The multi-view *Leo* sequence was captured at University of Bristol

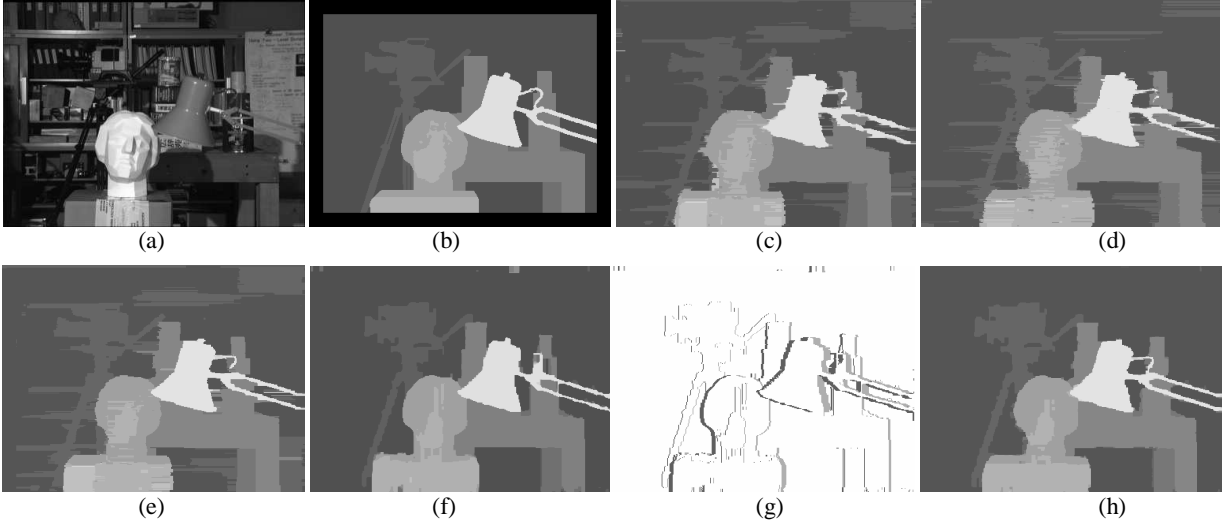


Fig. 2. (a) The middle view of the original image. (b) The true disparity. The horizontally estimated disparity map of *Head* multi-view images with (c) no any enhanced scheme. (d) variable window size but not shiftable window. (e) shiftable window (SAD) but not variable window size. (f) The vertically estimated disparity map with variable window size and shiftable window. (g) The occlusion map. The black and grey colours indicate the area that might be occluded in Left view reference and Right view reference respectively, whilst the white colour shows areas that could be seen in all views. (h) The estimated disparity map with vertical and horizontal combination.

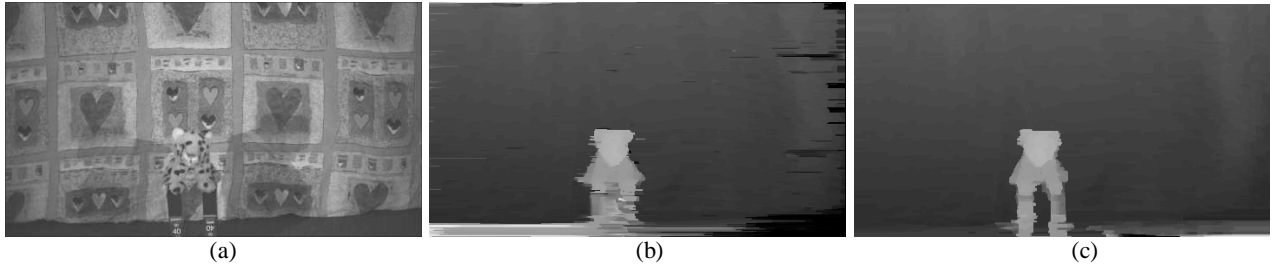


Fig. 3. The middle view of the *Leo* multi-view test images. (a) Original images. (b) Estimated depth map by dynamic programming without any enhance scheme. (c) Estimated depth map by dynamic programming with adaptive window size and shiftable window and also exploiting vertical and horizontal congregation.

Functions for non-parallel geometry, and the result is greatly improved when the adaptive window size, the shiftable window and the combination of vertical and horizontal scanning method are included as shown in Fig. 3 (c).

5. CONCLUSION

In this paper, a novel disparity/depth estimation based on window-based dynamic programming is presented and the cost functions to take into account discontinuity and occlusion are defined. The adaptive window size and shiftable window are included to improve the reliability and sharpness. According to the result, the proposed method provides more realistic disparity/depth map for all camera configurations. Moreover, the proposed horizontal and vertical scanning combination can highly improve the estimated disparity/depth map, if the two-dimensional array of multi-view camera allocation is available.

7. REFERENCES

1. Y. Ohta and T. Kanade, "Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 7, March 1985, pp. 139-154.
2. I. Cox, "A maximum likelihood N-Camera stereo algorithm," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 733-739.
3. T. Kanade, M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," IEEE Trans. on Pattern Anal. and Mach. Intell., 1994, pp. 920 – 932.
4. I. Cox, S. Hingorani, and S. Rao, "A Maximum Likelihood Stereo Algorithm," Computer Vision and Image Understanding, Vol. 63, No. 3, May 1996, pp. 542–567.
5. Peter N. Belhumeur, David Mumford, "A Bayesian Treatment of the Stereo Correspondence Problem Using Half-Occluded Regions," Proceedings of CVPR '92, 15-18 June 1992, pp. 506 – 512.
6. N. Grammalidis and M. Strintzis, "Disparity and Occlusion Estimation in Multiocular Systems and Their Coding for the Communication of Multiview Image Sequences," IEEE Trans. on Circuits and System for Video Technology, Vol.8, No.3, Jun 1998, pp. 328-344.
7. R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," 2nd edition, Cambridge University Press, 2004.
8. Y. Ohta and T. Kanade, "Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 7, March 1985, pp. 139-154.