Cortical Multiscale Line-Edge Disparity Model

J.M.F. Rodrigues, J.A. Martins, R. Lam, and J.M.H. du Buf

Vision Laboratory, LARSyS, University of the Algarve, 8005-139 Faro, Portugal
{jrodrig,jamartins,rlam,dubuf}@ualg.pt

Abstract. Most biological approaches to disparity extraction rely on the disparity energy model (DEM). In this paper we present an alternative approach which can complement the DEM model. This approach is based on the multiscale coding of lines and edges, because surface structures are composed of lines and edges and contours of objects often cause edges against their background. We show that the line/edge approach can be used to create a 3D wireframe representation of a scene and the objects therein. It can also significantly improve the accuracy of the DEM model, such that our biological models can compete with some state-of-the-art algorithms from computer vision.

1 Introduction

Developing better biological models to extract scene disparity is important for applications like cognitive robotics, as the relative positions and shapes of 3D objects in the environment are important for navigation and object recognition. Depth information allows to separate occluding scene components, i.e., objects which are in front of other objects.

Although we do not exactly understand how the visual system extracts disparity, this process occurs after the lateral geniculate nuclei which relay information from the left and right retinae to the primary area V1 of the visual cortex, in the cortical hypercolumns [3]. This is the first cortical processing stage where disparity can be prepared for subsequent processing in many other areas devoted to motor control, focus-of-attention, object segregation and recognition with partial occlusions.

In computer vision there are numerous approaches for stereo vision [10,12], but only few are biologically motivated. As for one of the most recent biological models [5], most have one common aspect: they are based on the widely applied disparity energy model (DEM). One of the exceptions is the model by Pugeault et al. [6], which combines geometric information with multi-modal constraints of local line/edge features. In a previous paper [4] we presented an improved version of the DEM model and obtained good results when applying it to real-world images (ranked as "BioDEM" on [1]). Two problems of our DEM model must be solved before it can seriously compete with computer vision algorithms [9,10,11], namely the lack of precision at boundaries and in large untextured regions, i.e., larger than the receptive fields of cells used in the DEM model.

Here we present a new disparity model based on multiscale line and edge coding: the DLE (Disparity Line/Edge) model. Our main contributions are a

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part I, LNCS 7324, pp. 296–303, 2012.

biologically plausible disparity model that can extract and assign disparity to lines and edges, and we show that is possible to combine information from the DLE and DEM models such that we can compete with state-of-the-art computer vision algorithms. We also show that the DLE model can be easily used to build a 3D wire-frame model of real-world scenes.

Line and edge coding is explained in Section 2. Section 3 deals with the DLE model. The integration of the DLE and DEM models is presented in Section 4. In Section 5, experimental results are presented together with a ranking on [1]. We conclude with a small discussion in Section 6.

2 Multiscale Line and Edge Coding

In the cortical hypercolumns in area V1 [3], the columnar structure brings retinotopic projections from the left and right eyes closely together. Cells with large dendritic fields and synapses in neighbouring left-right columns can receive input from both eyes. One might therefore assume that first disparity processing occurs already in V1. But there is more: in V1 we find simple, complex and end-stopped cells, which are thought to play an important role in coding the visual input: to extract multiscale line/edge and keypoint information (keypoints are line/edge vertices or junctions, but also blobs). If lines and edges are extracted in area V1 where also left and right retinal projections are closely together, one might even assume that depth is attributed to them. In other words, that a 3D wire-frame representation is built in V1 for handling 3D objects and scenes. Although this idea is speculative, many V1 cells have been found to be tuned to different combinations of frequency (scale), orientation, colour and disparity. If not coded explicitly, disparity could be coded implicitly. In our DLE model we assume that lines, edges and disparity are coded explicitly.

Responses of even and odd simple cells, corresponding to the real and imaginary parts of a Gabor filter [8], are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, i being the orientation (we use 8). The scale s is given by λ, the wavelength of the Gabor filters, in pixels. We use $4 \leq \lambda \leq 24$ with $\Delta\lambda = 2$. Responses of complex cells are modelled by the modulus $C_{s,i}(x,y) = [\{R^E_{s,i}(x,y)\}^2 + \{R^O_{s,i}(x,y)\}^2]^{1/2}$.

The basic scheme for line and edge detection is based on responses of simple cells: a positive/negative line is detected where R^E shows a local maximum/minimum and R^O shows a zero crossing. In the case of edges the even and odd responses are swapped. This gives four possibilities for positive and negative events. An improved scheme [8] consists of combining responses of simple and complex cells, i.e., simple cells serve to detect positions and event types, whereas complex cells are used to increase the confidence. Lateral and cross-orientation inhibition are used to suppress spurious cell responses beyond line and edge terminations, and assemblies of grouping cells serve to improve event continuity in the case of curved events. We denote the line and edge map by $LE_s(x,y)$.

Keypoint maps are also exploited in the DLE model, as keypoints code line and edge crossings, singularities and points with large curvature. They are built from two types of end-stopped cells, single and double, which are the first and second derivatives of $C_{s,i}$. The latter are combined with tangential and radial

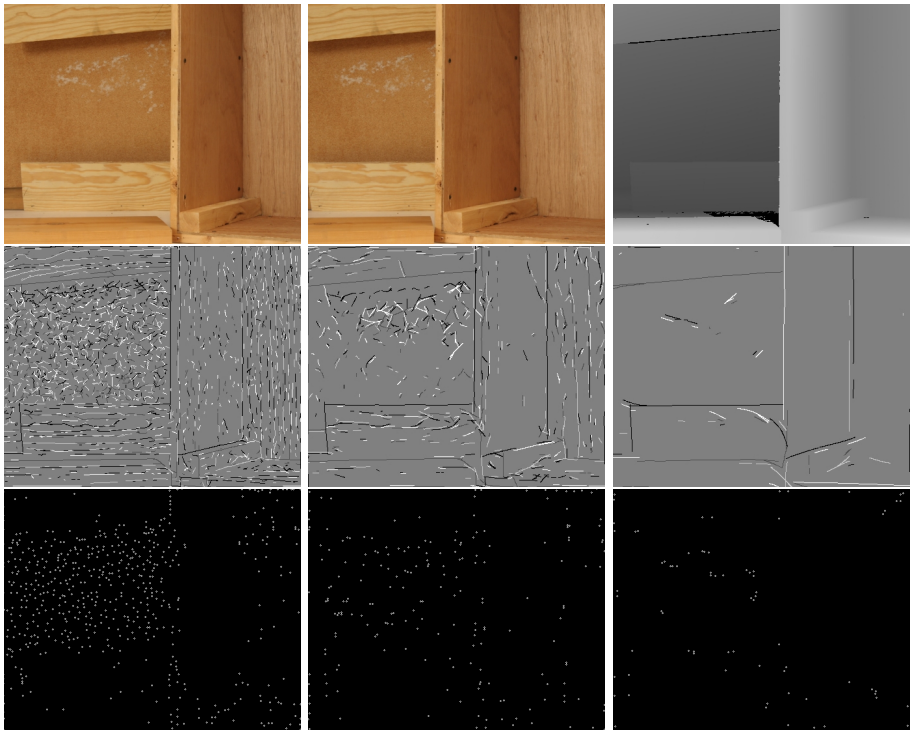


Fig. 1. Top row, left to right: a stereo pair (Wood1) and the ground truth of the left image from the Middlebury Stereo Datasets [10]. Middle row: multiscale line and edge coding of the left image at $\lambda = \{4, 16, 24\}$. Bottom row: detected keypoints (diamond symbols) at the same scales.

inhibition schemes in order to obtain precise keypoint maps $KP_s(x, y)$. For a detailed explanation with illustrations see [7].

The top row of Fig. 1 shows, left to right, a stereo pair (Wood1) and the disparity map of the right image from the Middlebury Stereo Datasets [1]. The middle row shows the multiscale line and edge coding at scales $\lambda = \{4, 16, 24\}$. Different gray levels, from white to black, show detected events: positive/negative lines and positive/negative edges, respectively. As can be seen, at fine scales many small events are detected, whereas at coarser scales only global structures remain. The bottom row shows detected keypoints (diamond symbols) at the same scales. The $LE_s(x, y)$ and $KP_s(x, y)$ maps will be exploited below.

3 Line and Edge Disparity Extraction and Assignment

The disparity to be assigned to each line/edge event (pixel) is based on the left-right correspondence over scales. First we suppress events which may be due to noise. At each scale s of the left (l) and right (r) image of the stereo pair $LE_{s,l/r}(x, y)$, we compute the maximum response of the monocular complex

cells $C_{s,i}$, but only at positions where events are detected. All events with a $C_{s,i}$ below 5% of the maximum response are inhibited. This yields $LEi_{s,l/r}(x,y)$.

In the right image, at each event position (x_r, y_r) and at the finest scale $(s = 1)$, $LEi_{1,r}$ is used to define regions of interest RoI which are centered on each (x_r, y_r). These RoI are grouping cells with a circular receptive field (RF). At the *same* positions (x_r, y_r), grouping cells with RFs are activated at all other scales, still in the right image, but the RF sizes are coupled to the scale: $2 \times \lambda_s$. This scale space of the right image, or hierarchical set of grouping cells with RFs at all event positions at $s = 1$, is used for accumulating displacement evidence of similar events at the same scales and relative (shifted) positions in the left image, $LEi_{s,l}$; see below. Basically, the RFs serve to "correlate" events in left and right images as a function of the shift. This is done at all individual scales, after which the scales are combined. The left scale space $LEi_{s,l}$ shifts in x (the epipolar lines) with $\delta x = 1$, such that $1 \leq \Delta x \leq 60$, hence a total of 60 shifts at which the events in both scale spaces are "correlated" in the RFs. The Δx with the maximum event correspondence is then assigned to the disparity map $D(x,y)$, where (x,y) corresponds to event positions (x_r, y_r) of $LEi_{1,r}$.

At each scale and within each RF, four correspondence measures are combined with different weight factors: (a) Counting all line/edge events with the same position, type (L/E) and polarity (+/–): $nLEtp_s$. (b) As (a) but only counting matching events irrespective of type and polarity: $nLEe_s$. (c) The number of complex cells with similar amplitudes at all event positions, i.e., $(C_{s,i,r} - 2) \leq C_{s,i,l} \leq (C_{s,i,r} + 2)$: $nLEa_s$. (d) Using $KP_{s,l/r}$, the number of keypoints with about the same coordinates, i.e., in small windows of size 3×3: nKP_s.

Before combining the four measures, they are first normalised: $nLEtp_s$, $nLEe_s$ and $nLEa_s$ are divided by the number of events in $LEi_{s,r}$, whereas nKP_s is divided by the number of keypoints in $KP_{s,r}$, each number being computed within each respective RF. The normalised numbers n are denoted by $\bar{n}$. The final correspondence is determined by combining the weighted and normalised measures over all scales: $\hat{C}_{\Delta x} = \sum_s (k_1 \times \bar{n}LEtp + k_2 \times \bar{n}LEe + k_3 \times \bar{n}LEa + k_4 \times \bar{n}KP)$, with $k_1 = k_4 = 4$ and $k_2 = k_3 = 1$ empirically determined weights (small changes do not change significantly the final result). Finally, the Δx with the maximum value of $\hat{C}$ is stored in the depth map $D(x,y)$. This map has an integer disparity resolution in pixels, like the ground-truth maps of real stereo pairs and the results of the DEM model. The maximum value of $\hat{C}$ can be seen as a confidence measure of the correspondence.

4 First Approach to DEM and DLE Integration

Our goal is to integrate the DLE and DEM models, and the DLE model has been explained above. In this section we briefly introduce the DEM model and the first integration with the DLE model.

The DEM [4] applies two neuronal populations: an encoding one that consists of a set of neurons tuned to a wide range of horizontal disparities, and a decoding one which exploits the responses of the encoding one for estimating local disparity. In the encoding population we use a set of 2880 binocular

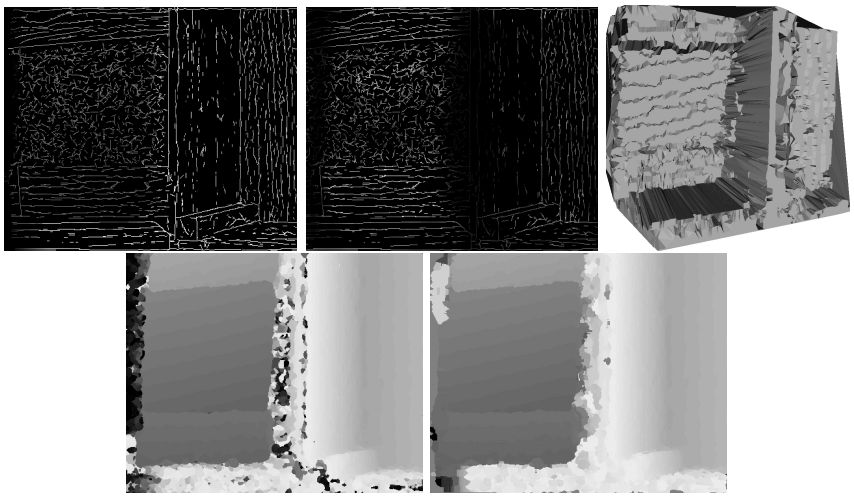


Fig. 2. Wood1 results. Top: DLE disparity map (left), certainty map (middle), and triangulated disparity map. Bottom: DEM map (left) and combined DLE-DEM result.

simple cells modelled by Gabor filters. These are used for building a total of 1440 phase-invariant binocular complex cells.

A population of binocular correlation detectors is used for initial disparity encoding. Normalising the stereo energy E to obtain the effective binocular correlation C removes the confounding effect of monocular contrast. This allows us to extract stimulus disparity from peaks in the population's activity code. We trained the population by exposing the cells to random-dot stereograms with known horizontal disparities. After the training phase, the same population is applied at all pixel positions (neighbourhoods) of real images. The disparity at each position is estimated by comparing the population code at that position with the previously learned codes. The disparity assigned at each pixel position is the disparity of the best-matching code by a winner-takes-all strategy. A detailed explanation can be found in [4].

The first DLE-DEM integration is rather elementary, but it serves our goal. For each event in the DLE map D we check the DEM map in a small region (event position plus its left, right, top and bottom neighbours). If it has about the same values ($\pm T_i$), where T_i is an integer threshold value, we copy the DEM cell responses of this region into D. If not, we fill the same region with the values of D. This way we fill open positions in D by creating a new disparity map $D_r(x, y)$. This process is applied recursively using the newly created $D_r(x, y)$ maps. If it is not possible to fill $D_r(x, y)$ anymore, but there are still gaps, we increase T_i and repeat the same procedure. In our experiments we increased T_i from 1 to a maximum of 5.

5 Experimental Results

We applied the models to five stereo pairs from the Middlebury Datasets. Figure 2 shows results obtained with the "Wood1" data (see Fig. 1). At top-left,

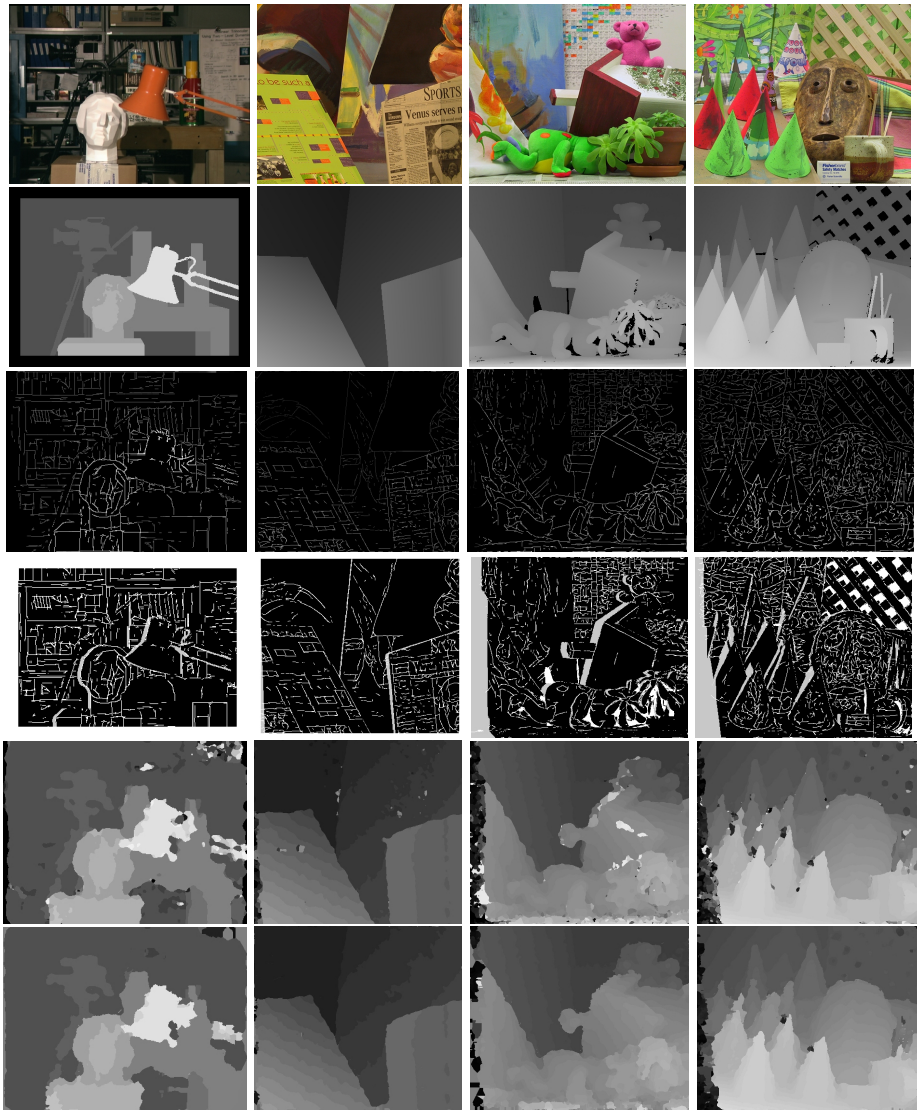
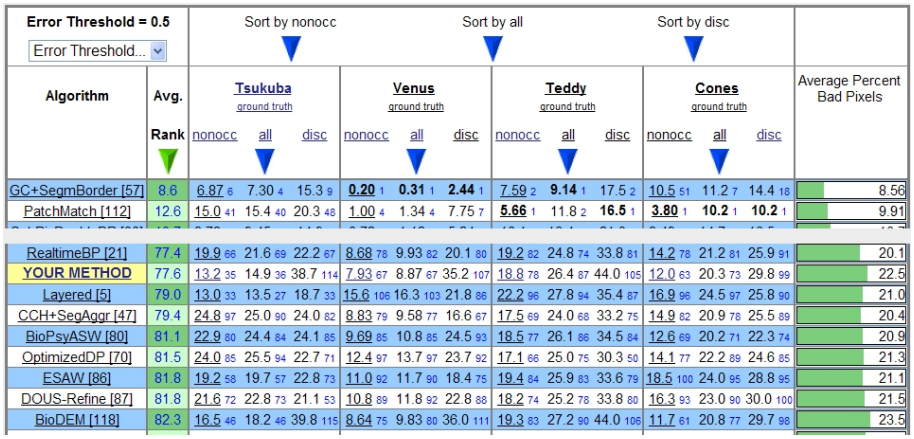


Fig. 3. Left to right: Tsukuba, Venus, Teddy and Cones data. Top to bottom: right images, ground truths, DLE results, DLE validated results, DEM maps, and combined DEM-DLE maps.

DLE's D map is coded by levels of gray, from dark (far) to white (near). The middle image shows the certainty of the matching process, where white is higher. We can see that the vertical "orthogonal" board has few events, therefore also with a lower certainty, because of its relatively untextured surface (Fig. 1). However, the board's narrow side has clear edges which are correctly represented in the disparity map, although their confidence is low because there are only two



Error Threshold = 0.5 (Error Threshold...)		Sort by nonocc			Sort by all			Sort by disc			
Algorithm	Avg. Rank	Tsukuba ground truth nonocc	Tsukuba all	Tsukuba disc	Venus ground truth nonocc	Venus all	Venus disc	Teddy ground truth nonocc	Teddy all	Teddy disc	Cones ground truth nonocc	Cones all	Cones disc	Average Percent Bad Pixels
GC+SegmBorder [57]	8.6	6.87 6	7.30 4	15.3 9	**0.20** 1	**0.31** 1	**2.44** 1	7.59 2	**9.14** 1	17.5 2	10.5 51	11.2 7	14.4 18	8.56
PatchMatch [112]	12.6	15.0 41	15.4 40	20.3 48	1.00 4	1.34 4	7.75 7	**5.66** 1	11.8 2	**16.5** 1	**3.80** 1	**10.2** 1	**10.2** 1	9.91
RealtimeBP [21]	77.4	19.9 66	21.6 69	22.2 67	8.68 78	9.93 82	20.1 80	19.2 82	24.8 74	33.8 81	14.2 78	21.2 81	25.9 91	20.1
YOUR METHOD	77.6	13.2 35	14.9 36	38.7 114	7.93 67	8.87 67	35.2 107	18.8 78	26.4 87	44.0 105	12.0 63	20.3 73	29.8 99	22.5
Layered [5]	79.0	13.0 33	13.5 27	18.7 33	15.6 106	16.3 103	21.8 86	22.2 96	27.8 94	35.4 87	16.9 96	24.5 97	25.8 90	21.0
CCH+SegAggr [47]	79.4	24.8 97	25.0 90	24.0 82	8.83 79	9.58 77	16.6 67	17.5 69	24.0 68	33.2 75	14.9 82	20.9 78	25.5 89	20.4
BioPsyASW [80]	81.1	22.9 80	24.4 84	24.1 85	9.69 85	10.8 85	24.5 93	18.5 77	26.1 86	34.5 84	12.6 69	20.2 71	22.3 74	20.9
OptimizedDP [70]	81.5	24.0 85	25.5 94	22.7 71	12.4 97	13.7 97	23.7 92	17.1 66	25.0 75	30.3 50	14.1 77	22.2 89	24.6 85	21.3
ESAW [86]	81.8	19.2 58	19.7 57	22.8 73	11.0 92	11.7 90	18.4 75	19.4 84	25.9 83	33.6 79	18.5 100	24.0 95	28.8 95	21.1
DOUS-Refine [87]	81.8	21.6 72	22.8 73	21.1 53	10.8 89	11.8 92	22.8 88	18.2 74	25.2 78	33.8 80	16.3 93	23.0 90	30.0 100	21.5
BioDEM [118]	82.3	16.5 46	18.2 46	39.8 115	8.64 75	9.83 80	36.0 111	19.3 83	27.2 90	44.0 106	11.7 61	20.8 77	29.7 98	23.5

Fig. 4. Benchmarking results [1]. Entry "BioDEM" is our improved DEM model. Entry "your method" is our result obtained by combining the DEM and DLE models.

vertical edges. The 3D wire-frame, top-right, was obtained by triangulating DLE's D map, using the Delaunay triangulation method of MeshLab [2]. Many errors in the DEM map (bottom-left) could be corrected by using the DLE map (bottom-right), although the disparity of the vertical board's side edges has not yet been fully exploited.

More results are shown in Fig. 3. Comparing the DEM results (2nd line from bottom) with the DEM-DLE combined ones (at bottom), many small regions and contours have been improved or corrected, although obtaining precise contours requires more advanced processing. Figure 4 shows benchmarking results with the lowest error threshold of 0.5 [1], which already included our improved DEM model [4] in the "BioDEM" entry. Combining DEM with DLE improved the overall ranking from 82.3 to 77.6 (the "your method" entry). If the table is sorted by evaluating only non-occluded regions (*nonocc*), our method jumps to the top half of the table, which so far covers a total of 118 models. Our method is not yet so good when evaluating regions near depth discontinuities (*disc*). This was expected, because the DEM model is not good at transitions and the DLE model is used to improve the DEM results without yet exploring precise edge information (see Fig. 2 and its discussion above). Finally, to the best of our knowledge, our method is ranked highest when compared with other biologically inspired methods, except of course when sorted by *disc*.

6 Discussion

We presented a biologically inspired model to extract and assign disparity to lines and edges, exploiting multiscale simple and complex cells of the visual cortex. The DLE model yields good results, and the integration with the DEM model improves results such that they can compete with state-of-the-art computer

vision models. We have also shown that the DLE model can be used to create a 3D wire-frame model of a scene.

Nevertheless, important improvements remain to be done. One is a better integration of the DLE and DEM results in which borders defined by edges are preserved. At the moment, only events in the receptive fields are matched, with increasing RF size at coarser scales, but there is no link between geometric structures at the different scales. In addition, as shown by Pugeault et al. [6], spatial structures can be linked both in 2D and in 3D by using constraints like good continuity. Finally, the models are only applied to gray-scale images, completely ignoring colour information in the matching process. All these improvements are already being investigated.

Acknowledgements. This work was partially supported by the Portuguese Foundation for Science and Technology (FCT) project PEst-OE/EEI/LA0009/2011, EC project NeuralDynamics NeFP7-ICT-2009-6 PN: 270247, FCT project Blavigator RIPD/ADA /109690/2009 and by FCT PhD grant to JAM SFRH/BD/44941/2008.

References

1. Middlebury stereo vision – evaluation page, `http://vision.middlebury.edu/stereo/eval/` (last retrieved January 2012)
2. Cignoni, P., Corsini, M., Ranzuglia, G.: Meshlab: an open-source 3D mesh processing system. In: ERCIM News, pp. 45–46 (2008)
3. Hubel, D.H.: Eye, Brain and Vision. Scientific American Library series, vol. 22. Scientific American Library, New York (1995)
4. Martins, J.A., Rodrigues, J.M.F., du Buf, J.M.H.: Disparity energy model using a trained neuronal population. In: Proc. IEEE Int. Symp. on Signal Proc. and Inform. Technology, pp. 261–266 (2011)
5. Mutti, F., Gini, G.: Bio-inspired disparity estimation system from energy neurons. In: Proc. IEEE Int. Conf. on Appl. Bionics and Biomechanics, pp. 1–6 (2010)
6. Pugeault, N., Woergoetter, F., Krueger, N.: Disambiguating multi-modal scene representations using perceptual grouping constraints. PLoS ONE 5(6), e10663 (2010)
7. Rodrigues, J., du Buf, J.M.H.: Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. BioSystems 86, 75–90 (2006)
8. Rodrigues, J., du Buf, J.M.H.: Multi-scale lines and edges in V1 and beyond: brightness, object categorization and recognition, and consciousness. BioSystems 95, 206–226 (2009)
9. Scharstein, D.: High-accuracy stereo depth maps using structured light. In: Proc. IEEE Comp. Soc. Conf. on Comp. Vision and Pattern Recogn., pp. 195–202 (2003)
10. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: Proc. IEEE Conf. on Comp. Vision and Pattern Recogn., pp. 1–8 (2007)
11. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. of Computer Vision 47, 7–42 (2002)
12. Szeliski, R.: Stereo correspondence. In: Computer Vision, pp. 467–503. Springer, London (2011)