Disparity Energy Model Using a Trained Neuronal Population

Jaime A. Martins
Institute for Systems and Robotics
Vision Laboratory (FCT)
University of the Algarve, Portugal
Email: jamartins@ualg.pt

J.M.F. Rodrigues
Institute for Systems and Robotics
Vision Laboratory (ISE)
University of the Algarve, Portugal
Email: jrodrig@ualg.pt

J.M.H. du Buf
Institute for Systems and Robotics
Vision Laboratory (FCT)
University of the Algarve, Portugal
Email: dubuf@ualg.pt

***Abstract*—Depth information using the biological Disparity Energy Model can be obtained by using a population of complex cells. This model explicitly involves cell parameters like their spatial frequency, orientation, binocular phase and position difference. However, this is a mathematical model. Our brain does not have access to such parameters, it can only exploit responses. Therefore, we use a new model for encoding disparity information implicitly by employing a trained binocular neuronal population. This model allows to decode disparity information in a way similar to how our visual system could have developed this ability, during evolution, in order to accurately estimate disparity of entire scenes.**

***Keywords*—disparity, population coding, learning, biological model**

I. Introduction

One of the intriguing functions of our visual cortex is to extract disparity information from the surrounding environment. This is done after the lateral geniculate nuclei (LGN), where information from the left and right retinae are relayed to the primary area V1, in the cortical hypercolumns [1]. This is the first cortical processing stage, but disparity extracted there plays an important role in many other areas devoted to motor control, from walking around to precise eye-hand coordination, focus-of-attention, and object segregation plus recognition with partial occlusions. The development of better models is important to deepen our insights, but also for many practical applications, like in robotics where the same issues arise. In computer vision there are numerous approaches for stereo vision [2], but only few are biologically motivated. As for one of the most recent biological models [3], most have one common aspect: they are based on the widely accepted Disparity Energy Model (DEM). It was first introduced from research into the cat's visual pathways and cortex [4].

Recent research into the composition of disparity energy neurons [3] has lead to different combinations of DEM subunits into an energy complex cell, with different weights and signs. A common goal is to better explain the disparity-tuning curves of neurons in the rhesus monkey [5]. The use of windowed cross-correlation between the left and right eye's images to measure disparity could explain some biological limits of stereopsis [6]: the *disparity-gradient limit* concerns the inability to perceive depth when the change in disparity within a region is too big, and *stereoresolution* is the inability to perceive spatial variations in disparity that occur at scales which are too fine. In the case of uniform-disparity random-dot stereograms, the DEM model was even able to explain that neurons tuned to horizontal disparities can also discriminate vertical disparities [7].

There have been many innovations, but the latest biological models have been applied to real-world scenes only very recently [8]. The main reason for this lag is that models are being tested by specific stimuli, such as random-dot stereograms or bar and grating patterns, in order to evaluate a model's theoretical performance [7], or to prepare psychophysical experiments with minimal random noise [5]. Used test patterns are far away from real-world and complex images. The latter require massive signal processing and still keep a good accuracy. To the best of our knowledge, this paper is one of the few in which a DEM model is applied to real images.

In our DEM implementation we use two neuronal populations for obtaining disparities: (1) An **encoding population** which consists of a set of neurons tuned to a wide range of parameters such as horizontal disparities, spatial frequencies and orientations; this is further explained in Section III-A. (2) A **decoding population**, with the same range of parameters, for estimating the disparity; this is further explained in Section III-B. We use an encoding method similar to that of Read [7], which is based on the DEM model [4], with proper normalization to yield a local correlation value with neighborhood weighting [6], [9]. The activity of the encoding population is subsequently decoded by a separate, higher-level population, using a template-matching process similar to that of [7], [10].

Our main contributions are the adaptation of the biologically plausible DEM model to separate encoding and decoding populations, the prior training of these populations, the extraction of disparity values in entire scenes, and the application to real-world images, with good results.

II. Disparity Energy Model

In cortical area V1 there are simple, complex and end-stopped cells (the latter we do not employ here). Monocular receptive fields (RFs) of simple cells can be modeled by Gabor wavelets [11], [12]. Their parameters specify the preferred orientation θ, spatial frequency f, receptive field size σ and spatial phase

978-1-4673-0753-6/11/$26.00 ©2011 IEEE

ϕ. Binocular cells can be based on pairs of simple cells with different RFs, such that they can signal disparity if a same but shifted pattern is present in the RFs. However, binocular simple cells do not reliably signal disparity because they are also sensitive to the contrast and position of the pattern within their fields: disparity-tuning curves of simple cells as measured with bright and dark bars, which have different Fourier phases, are very different [4]. The problem is that such tuning curves strongly depend on the Fourier phase of the pattern [11]: any change to a pattern other than an amplitude scaling (average brightness and contrast) alters the Fourier phase ϕ, which in turn affects disparity tuning. According to [11], if R_S is the response of a pair of simple cells with maximum amplitude ρ, Δx is a spatial shift and $\Delta\phi = \phi_L - \phi_R$, then

$$R_S \approx 2\rho \cos\left(\phi + \frac{\Delta\phi}{2} - \pi f \Delta x\right) \cos\left(\frac{\Delta\phi}{2} + \pi f \Delta x\right). \tag{1}$$

By contrast, *complex* cells do not have separate excitatory and inhibitory subregions within their receptive fields, so they are not sensitive to local phase (but still to position, orientation and size of a pattern). A phase-independent binocular complex cell can be made from two simple cells $s1$ and $s2$ provided that their phase difference $|\phi_{s1} - \phi_{s2}| = \pi/2$, i.e., they are in quadrature. The response of a complex cell is obtained by summing the squared responses of the two simple cells. According to [11] this yields the complex cell's response

$$R_C \approx 4\rho^2 \cos^2\left(\frac{\Delta\phi}{2} + \pi f \Delta x\right). \tag{2}$$

Complex cells are insensitive to contrast polarity within their RF and only broadly selective to stimulus position [4]. They have also been found to be sensitive to fine binocular disparity, and only complex cells respond to dynamic random-dot stereograms [13]. This class of stereograms maintains a constant disparity over time but the actual arrangement of the dots, and hence the Fourier phase, changes randomly from frame to frame. As simple cells are sensitive to the phase, they lose their disparity tuning as a result of averaging over the random phases of the dot patterns. Complex cells also have a much finer disparity selectivity than what would be predicted by the size of their RFs [4]. An important advantage of binocular complex cells is that they respond differently to inverted local pattern polarities at their preferred disparity, in contrast to monocular complex cells [4]. For this reason we first employ binocular simple cells and their responses are then combined by binocular complex cells.

We could estimate the preferred disparity D_{pref} of a binocular complex cell from its RF properties [11]: the phase difference $\Delta\phi$, spatial frequency f, orientation θ and the RF's position difference Δx. This yields

$$D_{\text{pref}} \approx \frac{\Delta\phi}{2\pi f \sin\theta} + \Delta x, \tag{3}$$

which means that the cell's response is maximal when the RF is stimulated by the preferred disparity. Unfortunately the brain cannot explicitly obtain D_{pref}, as it has no access to such intrinsic cell parameters, only cell responses.

III. Methods

A. Disparity encoding population

Similar to real binocular simple cells with both position and phase disparity [5], these cells are modeled by using two monocular RFs with the same size, orientation and spatial frequency, but with different phases ϕ and positions on the retinae Δx; see Eqns 4 and 5 and the detailed explanation in the next paragraph. During the training phase, all cells are located at the center of the fovea, i.e., the center of the RFs is at position $(0, 0)$). Since we use cells tuned to different orientations, the non-vertical ones will have horizontal and vertical phase disparity components. This problem is solved by introducing, for each cell orientation, a vertical position shift which compensates the phase component: $\Delta x_{\text{pos}} = \Delta x_{\text{enc}} - \Delta\phi \cdot (\cos(\theta)/2\pi f)$ and $\Delta y_{\text{pos}} = \Delta y_{\text{enc}} - \Delta\phi \cdot (\sin(\theta)/2\pi f)$, with Δx_{enc} and Δy_{enc} being the preferred horizontal and vertical disparities and subscript "enc" meaning encoding. We note that $\Delta y_{\text{enc}} = 0$ for all cells, as the vertical disparity of the cells in the fovea is expected to be zero, although it can be non-zero at other retinotopic positions [14].

The left (ρ_L) and right (ρ_R) RFs of the binocular simple cell are defined by

$$\rho_L(x, y; \theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) = \exp\left(-\frac{x_L'^2 + y_L'^2}{2\sigma^2}\right) \cos\left(2\pi f x_L' + \phi + \frac{\Delta\phi}{2}\right) \tag{4}$$

$$\rho_R(x, y; \theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) = \exp\left(-\frac{x_R'^2 + y_R'^2}{2\sigma^2}\right) \cos\left(2\pi f x_R' + \phi - \frac{\Delta\phi}{2}\right), \tag{5}$$

where x' and y' are the offset coordinates relative to the center $(0, 0)$ and rotated to the cell's preferred orientation:

$$x'_{L,R} = +\left(x \pm \frac{\Delta x_{\text{pos}}}{2}\right)\cos\theta + \left(y \pm \frac{\Delta y_{\text{pos}}}{2}\right)\sin\theta \tag{6}$$

$$y'_{L,R} = -\left(x \pm \frac{\Delta x_{\text{pos}}}{2}\right)\sin\theta + \left(y \pm \frac{\Delta y_{\text{pos}}}{2}\right)\cos\theta, \tag{7}$$

with $+$ signs for x'_L and y'_L, and $-$ signs for x'_R and y'_R, i.e., for crossed offsets.

For the encoding simple cells we selected a population set similar to [7], i.e.

- Horizontal position disparity Δx_{enc}: 60 values {0, ..., 59} in steps of 1 pixel.
- Orientation θ: 8 values $\{-67.5°, -45°, -22.5°, 0°, 22.5°, 45°, 67.5°, 90°\}$, where 90° is horizontal and 0° is vertical. Instead of applying only a few orientations, empirical tests showed that using more orientations yields better disparity estimates.
- Receptive field size σ: 3 values {2.8284, 2.0, 1.4142}. These are scaled by a factor of $\sqrt{2}$, as for the frequency (see below). Empirical results showed that bigger sizes

lead to too much blur at border regions, and smaller sizes introduce too much error in the case of the tested images.

- Spatial frequency f: 3 values $\{0.1768,\ 0.250,\ 0.3536\}$ cycles per pixel. These values are related to RF size by $\omega\sigma = \pi$ or $f = 1/2\sigma$. The frequency bandwidth at all scales was 1.14 octaves.
- Phase ϕ: 2 values $\{0, \pi/2\}$. Only two values are needed, $\pi/2$ apart, to build a phase-invariant complex cell from two simple cells in quadrature [4].
- Phase disparity $\Delta\phi$: 1 value $\{0\}$, implying no additional phase difference between the left and right RFs. Empirical tests showed that the use of phase differences (odd-symmetric disparity tuning curves) did not add significant information and sometimes even degraded the quality of disparity estimates; see also [3] and [5].

In total this selection yields a population of $60 \times 8 \times 3 \times 1 = 1440$ binocular complex cells (2880 simple cells).

1) Stereo energy model: The DEM model employs pairs of binocular simple cells in quadrature in order to construct phase-invariant complex cells. The responses of simple cells are obtained by the inner product (correlation instead of convolution) of each RF, left and right, and the corresponding image, left or right:

$$v_{L,R}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) = \iint \rho_{L,R}(x, y; \theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) \cdot I_{L,R}(x, y)\ dxdy, \quad (8)$$

where $I(x, y)$ is the input image with the average of all pixel values normalized to zero. In the standard energy model [4], the response of a binocular simple cell is $S = v_L^2 + v_R^2 + 2v_L v_R$, which can be split into the monocular term $M = v_L^2 + v_R^2$ and the binocular term $B = 2v_L v_R$.

For retrieving the local stereo energy E of a DEM complex cell which is invariant to the phases of local patterns in the input, it is necessary to sum the responses of binocular simple cells tuned to different phases,

$$E(\theta, f, \Delta\phi, \Delta x_{\text{enc}}) = \sum_{\phi_{1\to n}} \left[M(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) + B(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) \right]. \quad (9)$$

This stereo energy E represents something similar to the cross-correlation between the filtered and windowed images [6]. However, the value of E cannot be used directly as a disparity estimate, since its value not only reflects binocular energy (stimulus disparity between the left and right RFs), but also monocular energy (stimulus contrast inside each RF). This problem is solved by spatial pooling and effective binocular correlation as described next.

2) Spatial pooling: The RFs of real complex cells are larger than the modeled ones [11]. We exploit this property by averaging both terms (M and B), individually, over neighboring complex cells with overlapping RFs by using a Gaussian weighting function,

$$G_{sp} = k \exp\left(-(x^2 + y^2)/2\sigma^2\right), \quad (10)$$

with k a normalizing constant and σ the RF size. This yields M_{sp} and B_{sp}. This pooling operation involves simple grouping cells with dendritic field size defined by σ. This step stabilizes both values in case of real-world images with noise and non-uniform disparity ranges.

3) Effective binocular correlation: Our template matching is based on [7], using normalized correlation detectors [6], [9]. Based on the DEM, these detectors are normalized such that their response ranges between $+1$, when the left and right images are identical, and -1, when the left image is an inverted-contrast version of the right one. This is achieved by dividing the pooled binocular terms by the pooled monocular terms, after which C is also subjected to spatial pooling, similar as in 10, for robustness:

$$C(\theta, f, \Delta\phi, \Delta x_{\text{enc}}) = \frac{\sum_{\phi_{1\to n}} B_{sp}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}})}{\sum_{\phi_{1\to n}} M_{sp}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}})}. \quad (11)$$

This yields $C_{sp}(\theta, f, \Delta\phi, \Delta x_{\text{enc}})$.

Physiologically, this can be computed by combining the outputs of two energy neurons with phase disparities π apart. If such neurons are identical except for their phase disparities, then the first one computes $(M + B)$ and the second $(M - B)$. Both M and B are then available from the sum and difference of the two responses.

The quantity C relates to the correlation between local and filtered regions of the left and right eye's images [15], and takes values in the range of $[-1, 1]$. The population of binocular correlation detectors $C(\theta, f, \Delta\phi, \Delta x_{\text{enc}})$ is used for the initial encoding of disparity within the model. Recall that there are 10 different orientations, 6 different frequencies, and 60 different horizontal disparities, so the population consists of 1440 different correlation detectors.

Normalizing the stereo energy E to obtain the effective binocular correlation C removes the confounding effect of monocular contrast. This allows to extract stimulus disparity from peaks in the population's activity code. C has the useful property that it exactly equals 1 when the stimulus disparity matches the cell's preferred disparity. This holds for any pair of stereo images, irrespective of their spectral content etc., provided that the left eye's image is related to the right eye's one by exactly the same offset as that of the left and right receptive fields. If so, $v_L(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) = v_R(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}})$ for all $\theta, f, \phi, \Delta\phi$ and Δx_{enc}. Consequently, $2v_L v_R$ is the same as $v_L^2 + v_R^2$, and thus $C = 1$.

4) Model training: In this step we generate many examples of the population code to stimuli with known disparity. For this purpose we use random-dot stereograms with uniform disparity, generated by random values with a Gaussian distribution with zero mean and unit s.d., for a horizontal offset (Δx) between the left and right images. The gaps are filled by using randomly drawn pixels; see Fig. 1.

We trained the model to horizontal stimulation disparities Δx_{stim} ranging from 0 to 59 pixels with a stepsize of 1 pixel. For each disparity we generated 1000 random-dot pairs. Hence, training involved 60,000 stereograms.

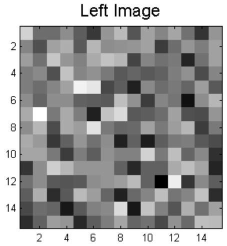
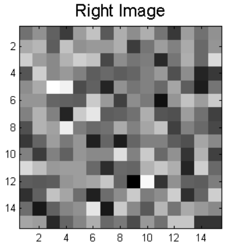


Fig. 1. Example of a 15×15 random-dot stereogram used in the training phase, with a uniform 2-pixel shift and thus horizontal disparity $\Delta x = 2$.

For each stereogram, the effective binocular correlation as described by Eq. (11) was computed. This parameter was then converted to a *mean spike count*, and averaged over the 1000 different stereograms. Averaging over random images serves to eliminate stimulus-dependent noise. This yields

$$W(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; \Delta x_{\text{stim}}) = \langle U[1 + C_{sp}(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; \Delta x_{\text{stim}})]\rangle . \quad (12)$$

Hence, W is the number of spikes produced by neurons tuned to orientation θ, frequency f, phase disparity $\Delta\phi$ and horizontal position disparity Δx_{enc}, averaged over all 1000 stimuli with the same disparity Δx_{stim}. In total, the trained population code consists of 1440 responses times 60 disparities. This training process, which is the core of the method, can be seen as a replication of visual learning in early childhood, assuming that basic neural circuitry is the result of evolution. Instead of training the population code at only one position, the center of the retina, it could be applied at all retinal positions with similar results.

B. Disparity decoding population

After training the encoding population, it is then applied to all pixel positions (neighborhoods) of real stereograms, excluding the border region. The disparity at each position is estimated by comparing the population code at that position with the learned codes. The disparity assigned to the position is the disparity of the best matching code [10]. Local disparity estimation is a simple matching process [10]: the input code of 1440 responses is matched or correlated with the 60 sets of 1440 trained codes. This is achieved by a hierarchy of subtraction and summation cells, the final output being selected by the winner-takes-all strategy. In reality this must be very fast, probably involving associative memory which can also be based on a training process [16].

Let $R_{\text{test}}(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; x, y)$ be the number of spikes fired by the encoding population at pixel position (x, y) of the test image. Remember that the population includes cells tuned to 8 orientations, 3 frequencies, 1 phase disparity and 60 horizontal disparities, so R_{test} is a set of 1440 spike counts at each image position. The local disparity is estimated by comparing $R_{\text{test}}(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; x, y)$ with the average spike counts W after training to the 60 stimulus disparities. For each possible disparity $(\Delta x_{\text{dec}}; x, y)$, where subscript "dec" means decoding, the correlation coefficient is calculated: $r(\Delta x_{\text{dec}}; x, y)$ is the correlation between the 1440 spike counts at position (x, y), $R_{\text{test}}(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; x, y)$, and the 1440 × 60 spike counts of $W(\theta, f, \Delta\phi, \Delta x_{\text{enc}}; \Delta x_{\text{dec}})$, so $r(\Delta x_{\text{dec}}; x, y) = \text{Corr}\left[(R_{\text{test}}(\xi; x, y), W(\xi; \Delta x_{\text{dec}})\right]$, where $\xi = \{\theta, f, \Delta\phi, \Delta x_{\text{enc}}\}$.

Mathematically, in the implemented matching process, the function $\text{Corr}(a, b)$ resembles the Pearson product-moment correlation coefficient between a and b:

$$r(\Delta x_{\text{dec}}; x, y) = \frac{\sum \left[\left\langle R_{\text{test}}^{(x,y)} W\right\rangle - \left\langle R_{\text{test}}^{(x,y)}\right\rangle \langle W\rangle\right]}{\sigma_{R_{\text{test}}^{(x,y)}} \cdot \sigma_W}, \quad (13)$$

where the sum, averages $\langle\rangle$ and standard deviations σ are taken over all ξ for each Δx_{dec} and (x, y). To avoid the problem of disparity in anti-correlated stereograms [17], we set any negative correlations to zero,

$$P(\Delta x_{\text{dec}}; x, y) = \lfloor r(\Delta x_{\text{dec}}; x, y)\rfloor \quad (14)$$

using halfwave rectification: $\lfloor x\rfloor = x$ for $x > 0$ and zero otherwise. Finally, the disparity assigned to position (x, y) is the value of Δx_{dec} with the maximum $P(\Delta x_{\text{dec}}; x, y)$.

We emphasize that no further processing is applied, i.e., the pixels' disparity values are not corrected using any continuity constraints in homogenious regions in combination with the detection of region boundaries.

IV. Results

We tested our method on various datasets, including the widely used stereograms *tsukuba*, *venus*, *teddy* and *cones* of the Middlebury stereo evaluation set [18], [19], also *aloe* and *cloth3* of the 2006 dataset, and *dolls*, *moebius* and *reindeer* of the 2005 dataset [20]. Fig. 2 shows all image pairs along with their groundtruth and our result.

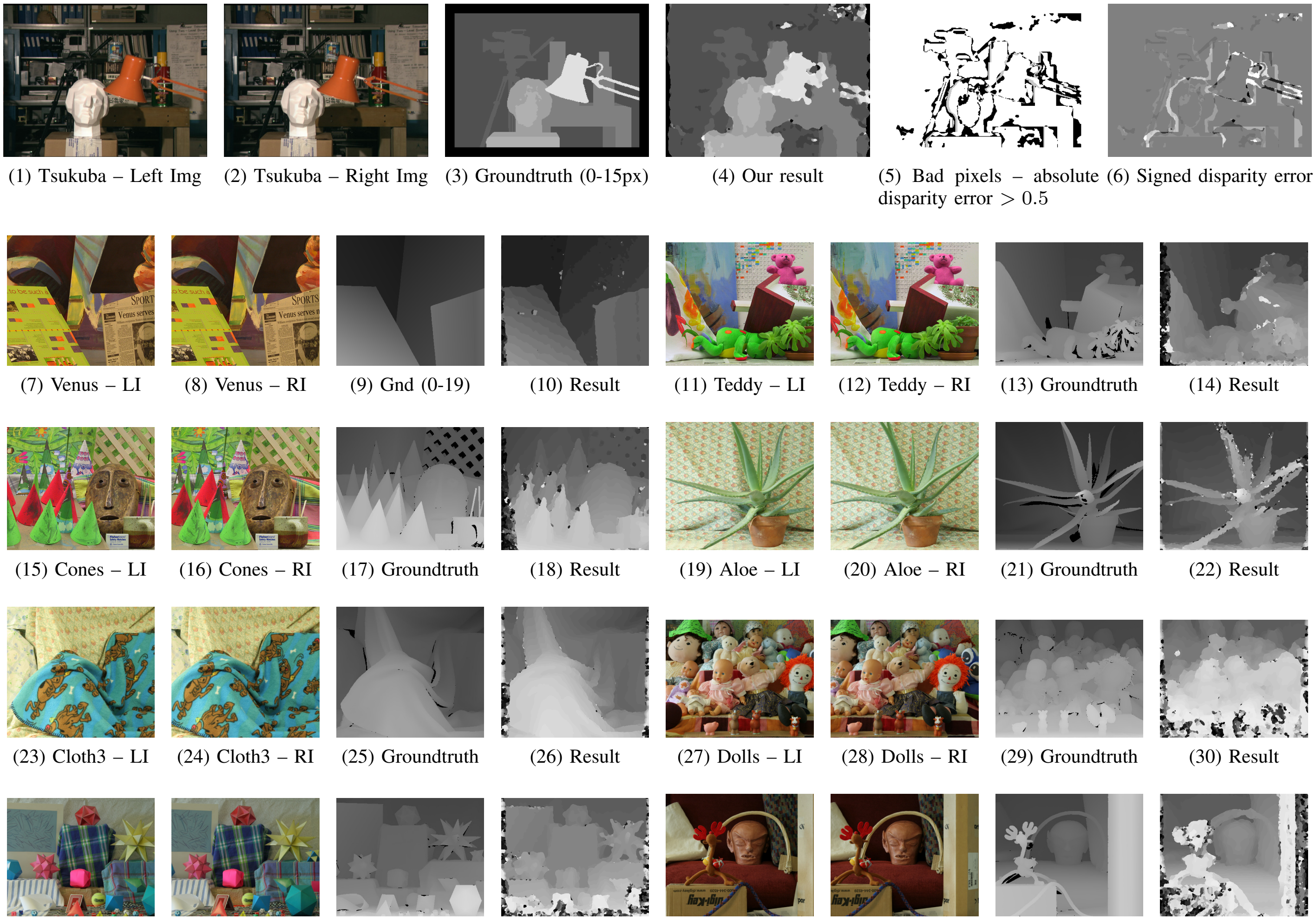
The algorithm obtained good results in the Middlebury evaluation test (see Fig. 3) [18], [19]. Best results were obtained for images without many small details. This is related to the size of the RFs in the cell population; smaller RFs are required to resolve smallest details.

Fig. 3 shows our result in part of the ranked results of other methods (which can include sophisticated postprocessing). This table was copied from the Middlebury online evaluation webpage. We applied the smallest available error threshold to emphasize that a biologically-inspired algorithm can achieve competitive results. Overall, we achieved a good position in the ranking table: rank 77.2 between 7.8 (best) and 108.1 (worst). Almost all methods ranked are from computer vision, most including postprocessing, which is not (yet) applied in our method.

V. Discussion and Conclusions

We presented an algorithm for disparity estimation based on the Disparity Energy Model. Unlike other models, it does not rely on filter parameters, it only employs filter responses. Therefore, the model must be trained, but only once. First, a population of binocular complex cells based on simple cells is defined. This population is trained by using random-dot stereograms. It is then applied at all image positions in order to compare its activation code with the learned activation code, resulting in local disparity estimates. Results obtained with the Middlebury evaluation database are quite good, taking into account that no sophisticated postprocessing has been applied. Our results are already better than those obtained by other promising bio-inspired algorithms [8].

The number of filters involved is rather large: 2880 simple cells on the basis of which 1440 complex cells are constructed, all this at every retinotopic (image) position. In reality, our visual cortex counts many more cells in area V1. However, our



(1) Tsukuba – Left Img (2) Tsukuba – Right Img (3) Groundtruth (0-15px) (4) Our result (5) Bad pixels – absolute disparity error > 0.5 (6) Signed disparity error

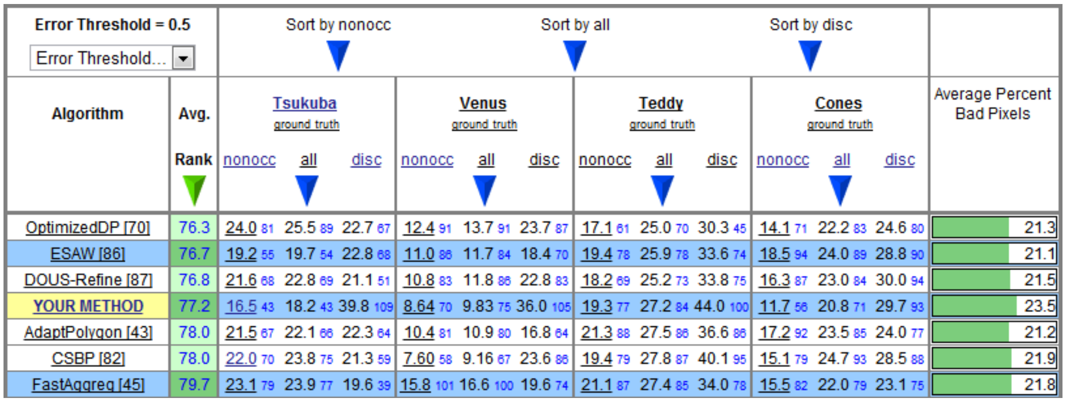
(7) Venus – LI (8) Venus – RI (9) Gnd (0-19) (10) Result (11) Teddy – LI (12) Teddy – RI (13) Groundtruth (14) Result

(15) Cones – LI (16) Cones – RI (17) Groundtruth (18) Result (19) Aloe – LI (20) Aloe – RI (21) Groundtruth (22) Result

(23) Cloth3 – LI (24) Cloth3 – RI (25) Groundtruth (26) Result (27) Dolls – LI (28) Dolls – RI (29) Groundtruth (30) Result

(31) Moebius – LI (32) Moebius – RI (33) Groundtruth (34) Result (35) Reindeer – LI (36) Reindeer – RI (37) Groundtruth (38) Result

Fig. 2. Results obtained with the Middlebury stereo dataset [18], [19]. The groundtruth images mention in parenthesis the complete range of disparity values if it differs from the default of 0-59 pixels.



Error Threshold = 0.5 Error Threshold... ▾		Sort by nonocc			Sort by all			Sort by disc						
Algorithm	Avg. Rank	Tsukuba ground truth			Venus ground truth			Teddy ground truth			Cones ground truth			Average Percent Bad Pixels
		nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	
OptimizedDP [70]	76.3	24.0 81	25.5 89	22.7 67	12.4 91	13.7 91	23.7 87	17.1 61	25.0 70	30.3 45	14.1 71	22.2 83	24.6 80	21.3
ESAW [86]	76.7	19.2 55	19.7 54	22.8 68	11.0 86	11.7 84	18.4 70	19.4 78	25.9 78	33.6 74	18.5 94	24.0 89	28.8 90	21.1
DOUS-Refine [87]	76.8	21.6 68	22.8 69	21.1 51	10.8 83	11.8 86	22.8 83	18.2 69	25.2 73	33.8 75	16.3 87	23.0 84	30.0 94	21.5
YOUR METHOD	77.2	16.5 43	18.2 43	39.8 109	8.64 70	9.83 75	36.0 105	19.3 77	27.2 84	44.0 100	11.7 56	20.8 71	29.7 93	23.5
AdaptPolygon [43]	78.0	21.5 67	22.1 66	22.3 64	10.4 81	10.9 80	16.8 64	21.3 88	27.5 86	36.6 86	17.2 92	23.5 85	24.0 77	21.2
CSBP [82]	78.0	22.0 70	23.8 75	21.3 59	7.60 58	9.16 67	23.6 86	19.4 79	27.8 87	40.1 95	15.1 79	24.7 93	28.5 88	21.9
FastAggreg [45]	79.7	23.1 79	23.9 77	19.6 39	15.8 101	16.6 100	19.6 74	21.1 87	27.4 85	34.0 78	15.5 82	22.0 79	23.1 75	21.8

Fig. 3. Middlebury evaluation dataset results [18], [19], with the smallest error threshold (0.5), ordered by average ranking. The table was extracted from the online evaluation webpage [21]. Obviously, the line “your method” should read “our method.”

978-1-4673-0753-6/11/$26.00 ©2011 IEEE

model does not introduce any new filters or many other cells, it only exploits the cells which are already available: pairs of simple cells at different positions are only wired together, and they also serve other purposes, like multi-scale line and edge coding, necessary for object recognition and brightness perception [12]. In addition, disparity, as for optical flow, is very important for object segregation, supplementing surface features like color and texture.

Interestingly, the fact that disparity is extracted in the hypercolumns of V1, where left and right projections are close together and where also lines and edges are coded, suggests that our visual system may attribute depth to detected lines and edges already at that level. Hence, our brain could use a sort of wireframe representation as used in computer graphics to model solid objects, and employ this for 3D object recognition. Furthermore, postprocessing of local disparity estimates can be based on edge information: edges between homogeneous regions are often caused by occlusions, exactly where disparity is not continuous and detail is visible in one projection but not in the other. Therefore, disparity estimation astride edges can be steered by detected edges, using phase tuning, and in homogeneous regions it can be smoothed.

Disparity training is applied to the encoding population in order to prepare the matching process, but the decoding population is a fixed neural network. It involves subtractive and divisive normalization in combination with halfwave rectification, for which plausible neuronal mechanisms have been proposed [7]. However, the decoding population could also be trained, even dynamically adapting itself to local image content by neural plasticity.

Recent research into the composition of disparity energy neurons [3], [5] suggests the use of a different combination of DEM subunits in an energy complex cell, with different weight and sign. This better explains disparity tuning curves of neurons in rhesus monkeys. It has been shown that the use of color information can enhance disparity maps, but with little biological background [22]. It would be interesting to train our model to color stereograms and test if results improve, especially in regions without textures where often wrong disparity estimates are generated because binocular RFs may have similar responses to a wide range of disparities; see e.g. the *tsukuba* image in Fig. 2 (4, the top-right corner). Also, disparity could be combined with color conspicuity around border regions [23] to better define disparity transitions.

Finally, coarse-to-fine-scale disparity estimation is a process which also deserves further attention, as its application has shown good results [11], with a very strong biological foundation [12].

Acknowledgments

This work was supported by the Portuguese Foundation for Science and Technology (FCT) through the PIDDAC Program funds (ISR/IST plurianual funding), EC project NeuroDynamics (FP7-ICT-2009-6 PN: 270247), FCT project Blavigator (RIPD/ADA/109690/2009) and FCT PhD Grant to Jaime A. Martins (SFRH-BD-44941-2008).

References

[1] D. H. Hubel, *Eye, Brain and Vision*, ser. Scientific American Library series. New York: Scientific American Library, 1995, vol. 22.

[2] R. Szeliski, "Stereo correspondence," in *Computer Vision*, ser. Texts in Computer Science, D. Gries and F. B. Schneider, Eds. Springer London, 2011, pp. 467–503.

[3] R. M. Haefner and B. G. Cumming, "Adaptation to natural binocular disparities in primate V1 explained by a generalized energy model," *Neuron*, vol. 57, pp. 147–158, 2008.

[4] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Encoding of binocular disparity by complex cells in the cat's visual cortex," *J. Neurophysiol.*, vol. 77, pp. 2879–2909, 1997.

[5] S. Tanabe and B. G. Cumming, "Mechanisms underlying the transformation of disparity signals from V1 to V2 in the macaque," *J Neurosc*, vol. 28, no. 44, pp. 11 304–11 314, 2008.

[6] H. R. Filippini and M. S. Banks, "Limits of stereopsis explained by local cross-correlation," *J. Vis.*, vol. 9, no. 8, pp. 1–18, 2009.

[7] J. C. A. Read, "Vertical binocular disparity is encoded implicitly within a model neuronal population tuned to horizontal disparity and orientation," *PLoS Comput. Biol.*, vol. 6, no. 4, p. e1000754, 2010.

[8] F. Mutti and G. Gini, "Bio-inspired disparity estimation system from energy neurons," in *1st IEEE International Conference on Applied Bionics and Biomechanics (ICABB-2010)*, oct. 2010.

[9] J. C. A. Read and B. G. Cumming, "Does visual perception require vertical disparity detectors?" *J Vision*, vol. 6, pp. 1323–1355, 2006.

[10] J. J. Tsai and J. D. Victor, "Reading a population code: a multi-scale neural model for representing binocular disparity," *Vision Research*, vol. 43, pp. 445–466, 2003.

[11] Y. Chen and N. Qian, "A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms," *Neural Computation*, vol. 16, no. 8, pp. 1545–1577, 2004.

[12] J. M. F. Rodrigues and J. M. H. du Buf, "Multi-scale lines and edges in V1 and beyond: Brightness, object categorization and recognition, and consciousness," *BioSystems*, vol. 95, pp. 206–226. doi:10.1016/j.biosystems.2008.10.006, 2009.

[13] G. F. Poggio, B. C. Motter, and Y. Squatrito, S.; Trotter, "Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random-dot stereograms," *Vision Res.*, vol. 25, pp. 397–406, 1985.

[14] J. C. A. Read, G. P. Phillipson, and A. Glennerster, "Latitude and longitude vertical disparity," *J Vision*, vol. 9, no. 13, pp. 11, 1–37, 2009.

[15] J. C. A. Read and B. G. Cumming, "Sensors for impossible stimuli may solve the stereo correspondence problem," *Nat Neurosci*, vol. 10, pp. 1322–1328, 2007.

[16] S. Yang and X. Yao, "Population-based incremental learning with associative memory for dynamic environments," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 5, pp. 542 –561, oct. 2008.

[17] J. C. A. Read and R. A. Eagle, "Reversed stereo depth and motion direction with anti-correlated stimuli," *Vision Research*, vol. 24, pp. 3345–3358, 2000.

[18] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

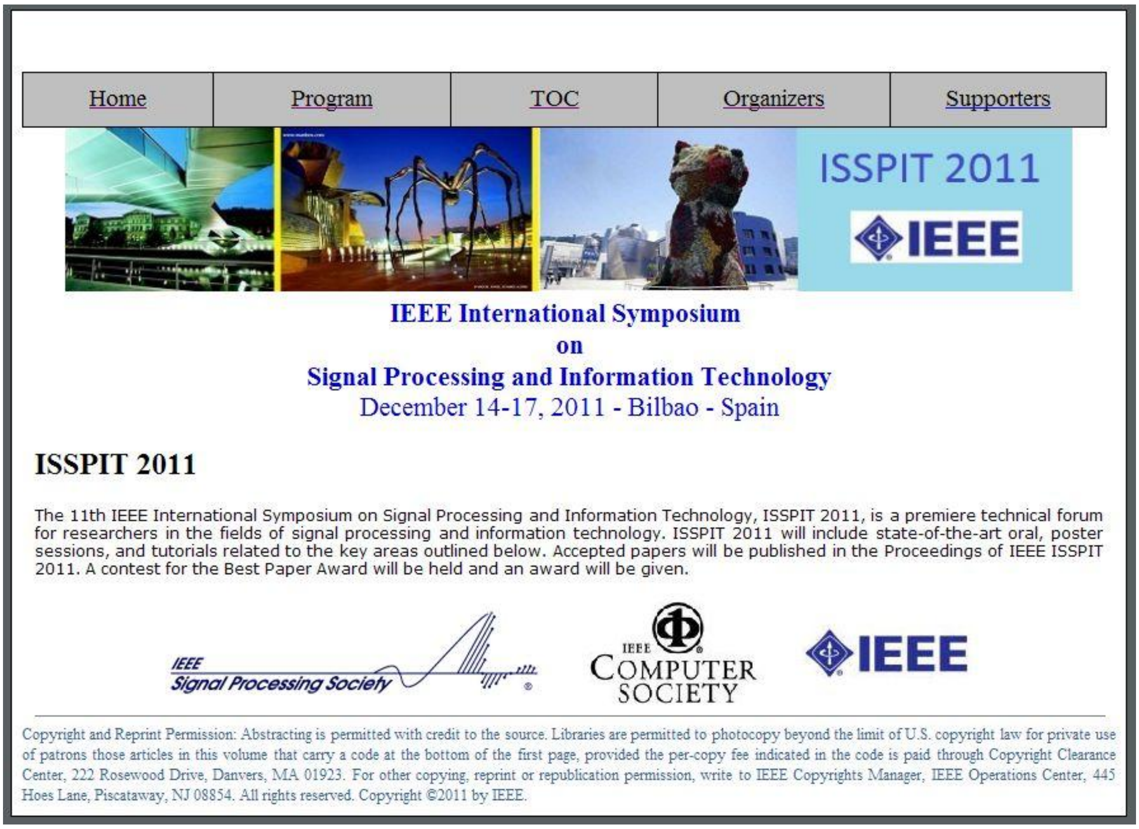
[19] D. Scharstein, "High-accuracy stereo depth maps using structured light," 2003, pp. 195–202.

[20] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[21] (Retrieved on 2011, July) Middlebury stereo vision page. [Online]. Available: http://vision.middlebury.edu/stereo/

[22] W. Miled and B. Pesquet-Popescu, "The use of color information in stereo vision processing," in *High-Quality Visual Experience*, ser. Signals and Communication Technology, M. Mrak, M. Grgic, and M. Kunt, Eds. Springer Berlin Heidelberg, 2010, pp. 311–330.

[23] J. A. Martins, J. M. F. Rodrigues, and J. M. H. du Buf, "Focus of attention and region segregation by low-level geometry," *Proc. Int. Conf. on Computer Vision Theory and Applications, Lisbon, Portugal*, vol. 2, pp. 267–272, 2009.



IEEE International Symposium
on
Signal Processing and Information Technology
December 14-17, 2011 - Bilbao - Spain

ISSPIT 2011

The 11th IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2011, is a premiere technical forum for researchers in the fields of signal processing and information technology. ISSPIT 2011 will include state-of-the-art oral, poster sessions, and tutorials related to the key areas outlined below. Accepted papers will be published in the Proceedings of IEEE ISSPIT 2011. A contest for the Best Paper Award will be held and an award will be given.