17^{th} Portuguese Conference on Pattern Recognition

Casa da Música, Porto
October 28, 2011

Editors: Jaime S. Cardoso, Luis F. Teixeira, Pedro Quelhas

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
FCT Fundação para a Ciência e a Tecnologia MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
INESCPORTO INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES DO PORTO LABORATÓRIO ASSOCIADO
INEB INSTITUTO DE ENGENHARIA BIOMÉDICA

Disparity energy model with keypoint disparity validation

Miguel Farrajota
elsio_farrajota@hotmail.com
Jaime A. Martins
jamartins@ualg.pt
J.M.F. Rodrigues
jrodrig@ualg.pt
J.M.H. du Buf
dubuf@ualg.pt

Institute for Systems and Robotics, Vision Laboratory,
University of the Algarve (FCT and ISE),
Campus de Gambelas, 8005-139 Faro, Portugal

Abstract

A biological disparity energy model can estimate local depth information by using a population of V1 complex cells. Instead of applying an analytical model which explicitly involves cell parameters like spatial frequency, orientation, binocular phase and position difference, we developed a model which only involves the cells' responses, such that disparity can be extracted from a population code, using only a set of previously trained cells with random-dot stereograms of uniform disparity. Despite good results in smooth regions, the model needs complementary processing, notably at depth transitions. We therefore introduce a new model to extract disparity at keypoints such as edge junctions, line endings and points with large curvature. Responses of end-stopped cells serve to detect keypoints, and those of simple cells are used to detect orientations of their underlying line and edge structures. Annotated keypoints are then used in the left-right matching process, with a hierarchical, multi-scale tree structure and a saliency map to segregate disparity. By combining both models we can (re)define depth transitions and regions where the disparity energy model is less accurate.

1 Introduction

One of the intriguing functions of our visual cortex is to extract disparity information from our surrounding environment. This is done after the lateral geniculate nuclei (LGN), where information from the left and right retinae is relayed to the primary area V1, in the cortical hypercolumns [3]. This is the first cortical processing stage. The development of good models is important to deepen our insights, but also for many practical applications. In computer vision there are numerous approaches for stereo vision [10], but only few are biologically motivated [2, 6, 11].

In our implemented disparity energy model (DEM) we apply two neuronal populations: (1) an encoding population that consists of a set of neurons tuned to a wide range of horizontal disparities, spatial frequencies and orientations, and (2) a decoding population which exploits the responses of the encoding population for estimating the local disparity. We use an encoding method similar to that of Read [6], with proper normalization to yield a local correlation value with neighbourhood weighting [2]. The activity of the encoding population is then decoded by using a template-matching process similar to that of [11].

The DEM model yields good results in regions where the depth is continuous [4], but it is less reliable at depth transitions. We therefore introduce a second model to extract disparity at corners, edge junctions, line endings and points with large curvature. This model is adapted from a previous one devoted to optical flow [1], because the matching of singularities like junctions between successive frames and left-right frames is similar. Evidence for joint encoding of motion and disparity in the visual cortex has been found [5], especially in the dorsal area MT. Hence, motion and disparity processing can be integrated.

In this keypoint disparity model (KDM), responses of end-stopped cells are used to detect keypoints and those of simple cells are used to detect orientations of the underlying vertex structures. Annotated keypoints are combined in a hierarchical, multi-scale tree structure and a saliency map to segregate disparity into regions. As will be shown, the DEM and KDM results can be combined to improve disparity estimation, but the KDM model critically depends on surface patterns like textures.

2 Disparity energy model DEM

For the *encoding* population we use a set of 2880 binocular simple cells, with left (L) and right (R) receptive fields (RFs) modeled by Gabor filters: sixty values of horizontal disparity $\Delta x_{\text{enc}} \in \{0, ..., 59\}$, eight orientations $\theta \in \{-67.5^\circ, -45^\circ, -22.5^\circ, 0^\circ, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ\}$, three receptive field sizes $\sigma \in \{2.8284, 2.0, 1.4142\}$ and spatial frequencies $f \in \{0.1768, 0.250, 0.3536\}$, and two phases $\phi \in \{0, \pi/2\}$. These are used for building a total of 1440 phase-invariant binocular complex cells. Responses of *simple* cells are obtained by the inner product (correlation) of each RF, left and right, and the corresponding image, left or right, yielding $v_{L,R}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}})$. In the standard energy model, 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. However, the value of E cannot be used directly as a disparity estimate, since it not only reflects binocular energy (stimulus disparity between the left and right RFs), but also monocular energy (pattern contrast inside each RF). This problem is solved by using normalized correlation detectors [2].

Based on the DEM, these detectors are normalized such that their responses range 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 [6]. This is achieved by dividing the binocular terms by the monocular terms, for all phases: $C(\theta, f, \Delta\phi, \Delta x_{\text{enc}}) = \sum_{\phi_{1\to n}} B_{sp}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}}) / \sum_{\phi_{1\to n}} M_{sp}(\theta, f, \phi, \Delta\phi, \Delta x_{\text{enc}})$. Response C relates to the correlation between local and filtered regions of the left and right images [6]. The population of binocular correlation detectors $C(\theta, f, \Delta\phi, \Delta x_{\text{enc}})$ is used for the initial encoding. 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.

We trained the population code by exposing the cells to stimuli with known disparity: we used random-dot stereograms with uniform disparity, generated by random numbers with a Gaussian distribution with zero mean and unit s.d., for a horizontal offset (Δx) between the left and right images. We trained the model to horizontal disparities Δx_{stim} ranging from 0 to 59 pixels with a stepsize of 1. For each disparity we generated 1000 random-dot pairs. Hence, training involved 60,000 stereograms. For each stereogram, the effective binocular correlation C was computed. This parameter was then converted to a *mean spike count* $(W = 1 + C)$, and averaged over the 1000 different stereograms. Averaging over random images serves to eliminate stimulus-dependent noise and to stabilise W. 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.

After the training phase, the same encoding population is applied at all pixel positions (neighbourhoods) of real stereograms, excluding the border region. The disparity at each position is estimated by comparing the population code at that position $(W_{[x,y]})$ with the learned codes $(W_{[0\to 59]})$. The disparity assigned to the position is the disparity of the best-matching code. Local disparity estimation is a simple matching process [11]: the input code of the 1440 W responses at each (x, y) 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.

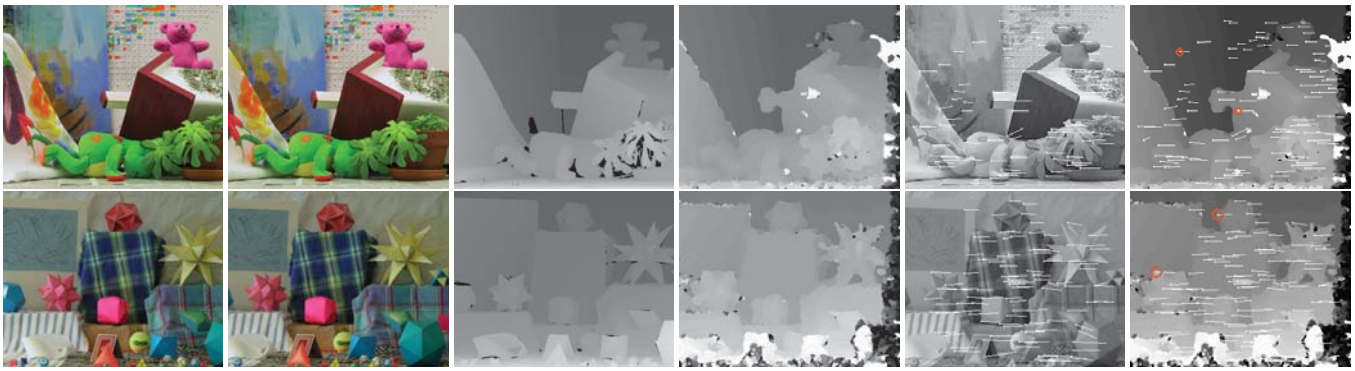

Figure 1: *Teddy* (top row) and *Moebius* (bottom row) images from the Middlebury dataset [9]. From left to right: left and right images, ground truth, results of the DEM model, results of the KDM model shown in the right image, and the combined KDM and DEM results.

3 Keypoint disparity model KDM

In the keypoint disparity model we apply keypoint detection by cortical end-stopped cells and optimized inhibition schemes as detailed in [8]. Here we use simple, complex and end-stopped cells at eight scales and with eight orientations. Again, RFs of simple cells are modeled by complex Gabor functions with sine and cosine components; complex cells by the modulus. The scale is given by λ, the wavelength of the simple cells in pixels: λ ={6, 9, 12,... 27}.

Detected keypoints are annotated by analyzing the responses of simple cells at three distances from their position, over orientation intervals around the 8 main orientations, taking only the orientations with the largest responses [1]. If $\Delta\varphi = 2\pi/8$, main orientations are $\varphi_k = k\Delta\varphi$ with $k = \{0,...,7\}$. With $\delta\varphi = \Delta\varphi/2$, the 8 orientation intervals are $\Phi_k = \varphi k \pm \delta\varphi$. The three distances are $\lambda/2$, λ and 2λ.

As in the case of optical flow [1], we use a multi-scale tree structure in which, at a very coarse scale, a root keypoint defines a single object, and at progressively finer scales more keypoints are linked which convey the object's details. However, at the coarsest scale applied, $\lambda = 27$, this may not be the case and an object may cause several keypoints. The tree structure links the keypoints over scales, from coarse to fine, with associated regions of influence at the finest scale. In order to determine which keypoints could belong to the same object we combine a saliency map with the multi-scale tree structure. We obtain the saliency map by summing responses of end-stopped cells over all scales. The latter, after thresholding, yields segregated regions which are intersected with the regions of influence of the tree. Therefore, the intersected regions link keypoints at the finest scale to segregated regions which are supposed to represent individual objects. Disparity is obtained by matching the annotated keypoints between right and left frames at all scales. This yields a displacement vector for all pairs of matching keypoints. Since we do not use superresolution, disparity is estimated in terms of integer pixel displacement, which facilitates the combination of DEM and KDM results.

4 Results

Fig. 1 shows two images from the Middlebury dataset [9] with the DEM and KDM results. The lighter DEM tones and longer KDM vectors represent the closest objects (highest disparity). It can be seen that the results of both models are good but not yet perfect. The red circles of a few keypoints have been introduced to indicate RF size of the scale applied. These illustrate that there are two situations in the RFs: (a) the DEM model yields a single disparity, in which case the KDM model can only validate the DEM result; or (b) the DEM model yields more than one disparity value, in which case the KDM model marks a depth transition and the DEM disparities in the keypoint's RF must somehow be corrected. This can be done by using continuity constraints, because the combined DEM and KDM disparities allow to detect fore- and background objects, and at transitions information in one frame may not be visible in the other one. This process can be extended by using edge information; see below.

5 Discussion

We presented two biological models for disparity estimation. The disparity energy model yields quite good results on the Middlebury evaluation database [4], taking into account that no sophisticated postprocessing is applied. However, the model lacks precision at depth transitions. The keypoint disparity model can improve precision at transitions, provided that keypoints are detected there. Keypoints are caused by surface patterns at the border between fore- and background objects, by the local shape and contrast. Experimental results showed that occasionally there are keypoints at transitions, but not many. Therefore, the DEM and KDM models are being complemented by yet another model: a multi-scale line- and edge-detection model [7] can be extended, like the keypoint model as shown here, to attribute disparity information to detected edges. In contrast to sparse keypoints at edge transitions, edges are more likely to occur there and they correspond to parts of a foreground object's contour.

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.

Acknowledgments

Supported by the Portuguese 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] M. Farrajota, J.M.F. Rodrigues, and J.M.H. du Buf. Optical flow by multi-scale annotated keypoints: A biological approach. *Proc. Int. Conf. on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2011), Rome, Italy, 26-29 January*, pages 307–315, 2011.

[2] S. Martin H. F. Banks. Limits of stereopsis explained by local cross-correlation. *J. Vis.*, 9(8):1–18, 2009.

[3] D. H. Hubel. *Eye, Brain and Vision*, volume 22. Scientific American Library, New York, 1995.

[4] J.A. Martins, J.M.F. Rodrigues, and J.M.H. du Buf. Disparity energy model using a trained neuronal population. *Subm. to IEEE Int. Symp. on Signal Proc. and Info. Technology (ISSPIT 2011), Bilbao, Spain, 14-17 Dec*, 2011.

[5] Peter Neri and Dennis M. Levi. Evidence for joint encoding of motion and disparity in human visual perception. *J Neurophysiol*, 100: 3117–3133, 2008.

[6] J. C. A. Read. Vertical binocular disparity is encoded implicitly within a model neuronal population tuned to horizontal disparity and orientation. *PLoS Comput. Biol.*, 6(4):e1000754, 2010.

[7] J. Rodrigues and J.M.H. du Buf. Visual cortex frontend: integrating lines, edges, keypoints and disparity. *Proc. Int. Conf. Image Anal. Recogn., Porto, Portugal*, Springer LNCS Vol. 3211:664–671, 2004.

[8] J.M.F. Rodrigues and J.M.H. du Buf. Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. *BioSystems*, 86:75–90, 2006.

[9] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.

[10] R. Szeliski. Stereo correspondence. In David Gries and Fred B. Schneider, editors, *Computer Vision*, Texts in Computer Science, pages 467–503. Springer London, 2011. ISBN 978-1-84882-935-0.

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