![HAL open science]

Patterns of Binocular Disparity for a Fixating Observer

Miles Hansard, Radu Horaud

Patterns of Binocular Disparity
for a Fixating Observer

Miles Hansard & Radu Horaud

INRIA Rhône-Alpes, 655 Avenue de l'Europe, Montbonnot 38330, France.
miles.hansard@inrialpes.fr, radu.horaud@inrialpes.fr

Abstract. Binocular information about the structure of a scene is contained in the relative positions of corresponding points in the two views. If the eyes rotate, in order to fixate a different target, then the disparity at a given image location is likely to change. Quite different disparities can be produced at the same location, as the eyes move from one fixation-point to the next. The pointwise variability of the disparity map is problematic for biological visual systems, in which stereopsis is based on simple, short-range mechanisms. It is argued here that the problem can be addressed in two ways; firstly by an appropriate representation of disparity, and secondly by learning the typical pattern of image correspondences. It is shown that the average spatial structure of the disparity field can be estimated, by integrating over a series of binocular fixations. An algorithm based on this idea is tested on natural images. Finally, it is shown how the average pattern of disparities could help to put the images into binocular correspondence.

1   Introduction

Binocular disparity is the difference in position of a matched point, as it appears in the left and right images. This difference can be divided into two components; one that is due to the structure of the scene, and one that is imposed by cameras themselves. In particular, the pattern of binocular disparity is sensitive to the relative orientation of the sensors. This is important for active vision systems, in which binocular fixation is achieved by rotating the cameras, such that the left and right images are centred on the point of interest. It follows that the pattern of disparity will be different for each fixation, even in a static scene.

The effect of relative orientation on disparity is problematic for biological visual systems, in which stereopsis is based on the output of local filter-like mechanisms [7]. For example, binocular cells in primate V1 have relatively small receptive fields, and may be tuned to a single direction of disparity on the retina. Hence it would be desirable for the visual system to arrange these mechanisms according to the patterns of disparity that occur most often. This would have two clear advantages. Firstly, depth-sensitivity could be improved, by placing additional mechanisms in regions of highly variable disparity. Secondly, when the image data are ambiguous, it would be useful to have an implicit model of the most likely disparity at each point on the retina. There is experimental

evidence to suggest that such an organization of disparity sensitivity exists in the primate visual cortex [1].

In section 2 it will be argued that the displacement of image-features between the left and right views is best represented with respect to the underlying epipolar geometry. However, as will be explained, the epipolar geometry depends on the relative orientation of the eyes. The question of whether the visual system uses this geometric information at the disparity-processing stage remains open. For this reason, we will consider both epipolar and non-epipolar representations in the present work.

If the visual system does estimate the epipolar geometry, then the pattern of disparities is largely determined by each fixation point; what remains to be estimated is the magnitude of each disparity. However, it is no less important to consider the average pattern of disparities in this case. The reason is simply that the same local mechanisms must be used during each fixation. Hence some arrangements of these mechanisms will be better than others, depending on which epipolar geometries are more likely to occur, as different points are fixated. This should make it clear that, although we use ideas from computer vision, the problems addressed here arise from *biological* constraints on visual processing. It should be emphasized that we are not directly investigating the distribution of scene depths [4] in this work. Rather, we are investigating the distribution of *disparity fields*, which is determined by the combination of eye-movements and scene structure.

Sections 2 and 3 describe the geometric and image-processing background that is subsequently required. Our main idea is presented in section 4, in which we show how a collection of disparity maps can be combined. This procedure is tested in section 5. A stereo image-pair is warped into a number of 'fixating' views, and the disparity field is recorded in each case. These maps are combined, to produce an average disparity map, with respect to the different fixation-points. We discuss, in section 6, how such a representation could be used by the visual system.

2 Disparity Models

The left and right eyes are modelled here by pinhole cameras, with centres of projection c_ℓ and c_r, respectively. It is convenient to represent image-points by their homogeneous coordinates $q_\ell = (x_\ell, y_\ell, 1)^\top$, and similarly for q_r. Suppose, without loss of generality, that the axes of the scene coordinate-system are aligned with the left eye. Then the image-points are related to the scene-point $\bar{q} = (x, y, z)^\top$ by the projections

$$z_\ell q_\ell = \bar{q} \quad \text{and} \quad z_r q_r = R(\bar{q} - c_r), \tag{1}$$

where R is the 3×3 rotation matrix that determines the relative orientation of the eyes. One possible representation of binocular disparity is simply the difference between q_ℓ and q_r. This is, in general, a vector with non-zero horizontal
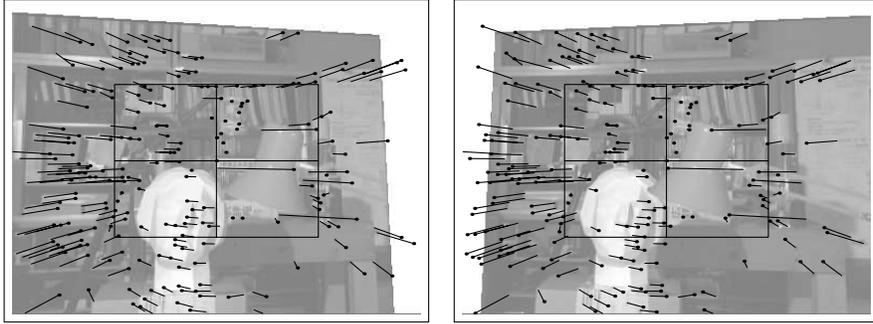
Fig. 1. A stereo image-pair that has been warped to simulate the fixation of a particular scene-point. The raw disparities of the matched points are indicated by the black vectors. The sharp-end of each vector marks the image feature; the blunt end marks the location of the same feature in the other view. No epipolar geometry has been imposed, and so the vectors do not follow a simple pattern. Images courtesy of the University of Tsukuba.

and vertical components [6]; we will call these the *raw* disparities. An example of a raw disparity field is shown in figure 1.

An alternative representation of disparity can be derived from the fact that the scene point $\bar{q}$ in equation (1) is equal to the back-projected image-point $z_\ell q_\ell$. It follows that the left and right image-points are related by the well-known equation

$$q_r \sim Rq_\ell + (1/z_\ell)e_r, \quad \text{where} \quad e_r = -Rc_r, \tag{2}$$

and '$\sim$' denotes equality up to a scalar multiple. The importance of this model is that if R is known, as well as q_ℓ and q_r, then only one degree of freedom, z_ℓ, remains for the unknown scene-point. The point e_r is the epipole, being the image of the left optical centre. Note that e_r varies with the relative orientation of the eyes, but not with the choice of scene-point $\bar{q}$. Another way to understand this is that the position of each point q_r is measured with respect to a corresponding reference-point Rq_ℓ in the same image. These reference points lie on the plane at infinity; however it can be shown that the same principles apply if the reference-points lie on any plane (not passing through either optical centre). This leads to the more general decomposition [2, 8];

$$q_r \sim Hq_\ell + \delta e_r, \tag{3}$$

in which H is a homography containing R and the parameters of the plane, while δ is proportional to the scalar depth of $\bar{q}$ with respect to the plane. The vector δe_r will be called the *epipolar* disparity, including equation (2) as a special case. We emphasize that the epipolar disparity has one degree of freedom δ, whereas the raw disparity has two; d_x and d_y. The epipolar disparity has several other advantages; for example, the reference plane can be chosen in order to reduce the size of the disparities. In our experiments, we use a fronto-parallel plane

(with respect to the head), positioned at the fixation distance. An example of the resulting epipolar disparity field is shown in figure 2.
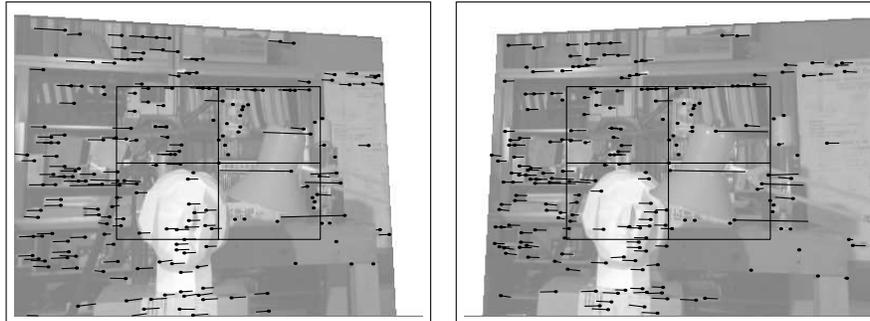

Fig. 2. The fixating stereo pair from figure 1 is shown again. The feature-correspondence is also identical, however, the epipolar geometry has been imposed. The blunt ends of the vectors now represent reference-positions on a virtual plane through the fixation-point. The disparities are organized along epipolar lines, and tend to be smaller than those in figure 1.

In order to recover the metric structure of the scene, it is necessary to know the relative orientation of the cameras, and to account for any geometric distortion imposed by the sensors. If these calibration parameters are unknown, then equation (3) can nonetheless be used to estimate non-metric properties of the scene. For example, it can be established whether a given scene-point is in front of or behind the reference plane encoded by $\boldsymbol{H}$. The effect of the fixation plane can be seen in figure 2. The plane is at a depth between that of the face (in the lower-left quadrant) and the far wall. The disparities associated with these two parts of the scene are in opposite directions. It has been argued elsewhere that a qualitative representation of this kind could explain several aspects of biological stereopsis [9].

3 Image Matching

In order to generate disparity-fields, we must have a stereo image pair with corresponding points identified. We use a simple feature-matching process, as follows. First we apply a Gaussian filter G, of width λ, to smooth each image I. We then construct an outer-product matrix from the luminance-gradient at each point. These matrices are themselves smoothed at scale μ, and the response $Q(x, y)$ is computed;

$$Q(x, y) = \frac{\det\big(G_\mu \star S\big)}{\operatorname{tr}\big(G_\mu \star S\big)}, \quad \text{where} \quad S(x, y) = \big(\nabla G_\lambda \star I\big)\big(\nabla G_\lambda \star I\big)^\top,$$

and '$\star$' denotes 2-D convolution. This commonly-used operator produces maxima in $Q(x, y)$ at 'interest points' $\boldsymbol{q}_{\ell i}$ and $\boldsymbol{q}_{rj}$ in the left and right images [3].

In order to match corresponding feature-points in the left and right views, we compare the colour of I_ℓ around $\boldsymbol{q}_{\ell i}$, to the colour of I_r around $\boldsymbol{q}_{rj}$. The 'cost' of matching these features is defined as the sum of squared colour-differences between the two image-patches $\mathcal{I}_\ell(\boldsymbol{q}_{\ell i})$ and $\mathcal{I}_r(\boldsymbol{q}_{rj})$;

$$F(\boldsymbol{q}_\ell, \boldsymbol{q}_r) = \frac{1}{\phi^2} \big| \mathcal{I}_\ell(\boldsymbol{q}_\ell) - \mathcal{I}_r(\boldsymbol{q}_r) \big|_{\mathcal{I}}^2, \tag{4}$$

where $| \cdot |_{\mathcal{I}}^2$ averages the pointwise squared-differences over the patches, and ϕ is a parameter relating to the expected photometric variation at corresponding points. The matching-costs are put into a table, F_{ij}, and the minima in each row i, and column j are computed:

$$m_{\ell i} = \arg\min_j \ F_{ij}, \quad \text{and} \quad m_{rj} = \arg\min_i \ F_{ij}.$$

We then enforce 'uniqueness' and 'compatibility' constraints on the matches, meaning that point $\boldsymbol{q}_{\ell i}$ matches $\boldsymbol{q}_{rj}$ if

$$i = m_{rj}, \quad j = m_{\ell i}, \quad \text{and} \quad F_{ij} < T\phi^2,$$

where $T\phi^2$ is a threshold defining the maximum photometric incompatibility between matched points. The procedure described above produces very sparse, but relatively reliable matches. Note that the matching cost in equation (4) does not penalize implausibly large disparities. The average pointwise magnitude of the disparity field is investigated below, and in section 6 it is shown how the resulting probabilistic model could be incorporated into the matching algorithm.

Our experimental data was constructed by applying appropriate homographies to an original stereo image pair, in order to simulate fixating pairs of views. In principle, we could apply the matching process to each pair of warped images. In practice, we compute the correspondence only once, using the original images. The homographies are then used to map the coordinates of the matched points into the fixating images. This is done in order to avoid irrelevant effects of the warping on the correspondence process; for example, pixel-resampling may reduce the number of points that are matched in more strongly warped images. We also enforce the epipolar constraint on the matched points, by considering only horizontal displacements in the rectified images.

4   Disparity Processing

In this section we describe our model of the disparity data. We have measured, in each image, the disparity of $k = 1 \cdots M$ points, over $v = 1 \cdots N$ fixations. Hence we have image positions $\boldsymbol{q}_{kv}$ and their associated (raw or epipolar) disparity vectors $\boldsymbol{d}_{kv}$. The procedures in this section apply to the left and right views independently, and so we suppress the ℓ, r indices, in order to simplify the notation.

We consider the data $\{\boldsymbol{q}_{kv},\,\boldsymbol{d}_{kv}\}$ as a single vector field, and ask what structure, if any, it contains. Note that the points $\boldsymbol{q}_{kv}$ are not evenly distributed in the images, and that neighbouring points may be associated with quite different disparities. Hence we effectively wish to smooth and interpolate the observed vector-field. We are particularly interested in the local-orientation of the field, and so the smoothing-process must treat vectors that differ in orientation by 180° as being 'similar'. This can be achieved by representing the disparities as outer-products

$$\boldsymbol{D}_{kv} = \boldsymbol{d}_{kv}\,\boldsymbol{d}_{kv}^{\top}, \tag{5}$$

each of which is a 2×2 matrix of rank-one [5]. As described above, we would like to have a representation of the average disparity at an arbitrary location $\boldsymbol{q}$, based on samples from points $\boldsymbol{q}_{kv}$. We use a simple kernel-like estimator to obtain

$$\boldsymbol{D}(\boldsymbol{q}) \propto \sum_{k}^{M} \sum_{v}^{N} W(\boldsymbol{q}_{kv}, \boldsymbol{q})\,\boldsymbol{D}_{kv}. \tag{6}$$

This gives the disparity-matrix $\boldsymbol{D}$ at position $\boldsymbol{q}$ as a weighted average over all of the data. The average is subsequently normalized by the sum of the weights. The kernel could be any decreasing function of the separation between $\boldsymbol{q}_{kv}$ and $\boldsymbol{q}$. We use an isotropic Gaussian, with width parameter w;

$$W(\boldsymbol{p}, \boldsymbol{q}) = \exp\left(\frac{-|\boldsymbol{p} - \boldsymbol{q}|^2}{2w^2}\right). \tag{7}$$

In general, the average matrices $\boldsymbol{D}(\boldsymbol{q})$ will have rank-two. The local orientation and variability of the disparity-field at location $\boldsymbol{q}$ is obtained by eigen-decomposition of the corresponding matrix. The eigenvector associated with the larger eigenvalue, σ_1^2, is oriented along the characteristic direction of disparity. The smaller eigenvalue, σ_2^2, indicates the variability of the disparity around the characteristic direction.

5 Simulation Results

In this section we investigate the distribution of raw and epipolar disparity fields by a simulation, based on real images. We believe that this approach is worthwhile, because it incorporates a number of effects that would be difficult to specify in a purely geometric simulation. For example, the joint distribution of feature-locations and scene-depths is naturally determined by the images themselves. Furthermore, it is possible to demonstrate the robustness of the smoothing process to the false matches contained in the disparity field. Data was generated by synthetically fixating each scene-point that had been matched in the images, and recording the resulting disparity field. As described in section 3, there was a single underlying set of correspondences; only the relative orientation of the two views was varied.

The procedure is complicated by the fact that the warped images are incomplete with respect to the original field of view (c.f. the edges of the images in

figures 1 & 2). The results would be biased if this effect were ignored, because it is the same structure (the upper and lower epipolar lines on the side of the epipole) which is lost in each case. We avoid this artifact by analyzing only the central 25% of the original field of view, defined by the inner rectangle in figures 1 & 2. We reject any fixation that would leave this region incomplete. The drawback of the approach is that the more variable disparities tend to lie in the periphery, and so our results are conservative.

The procedure described in section 3 returned 404 interest-points in the left image, and 398 in the right. Of these features, $M = 207$ were matched between the left and right images. There were 25 scene-points that could be fixated such that the resulting disparity-maps were complete over the central 25% of both images, for the reason described above. A further nine fixations were valid for the left image only, and a further one fixation for the right image only. All data was used in the analysis, meaning that several thousand disparities ($M \times N$, $M = 207$; $N = 26, 34$) contributed to each of the average disparity maps.

The distributions of disparity magnitude and orientation are shown in figures 3 and 4. As expected, the epipolar disparities are on average shorter than the original vectors; the means are 0.104 and 0.065, respectively. This difference is attributable to the use of a appropriate reference plane, as described in the introduction. It was also found that the epipolar lines were much less variable in orientation than the original disparity vectors; the standard deviations of the angular data are 0.448 and 0.168, respectively.
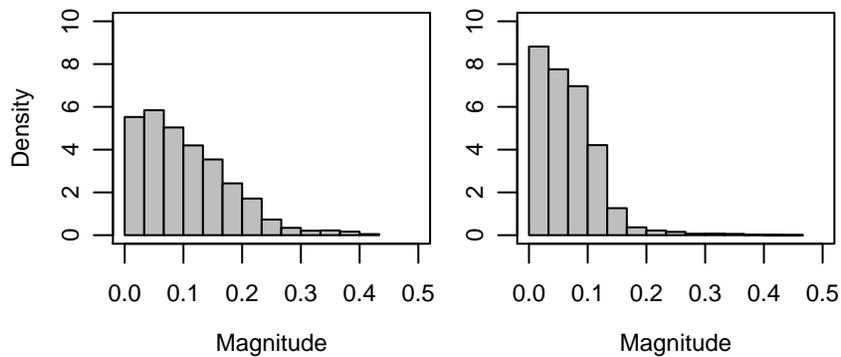

Fig. 3. Histograms of disparity magnitude for the raw (left) and epipolar (right) representations. The epipolar disparities are smaller, owing to the use of an appropriate reference plane.

Finally, we consider the spatial structure of the combined disparity maps. The estimator described in section 4 was used to resample the central region of the disparity maps on a regular grid, as shown in figure 5. The spatial width parameter w in equation (7) was set to one half of the grid spacing. It can be seen that raw disparity field is less regular than the epipolar field, as expected. The
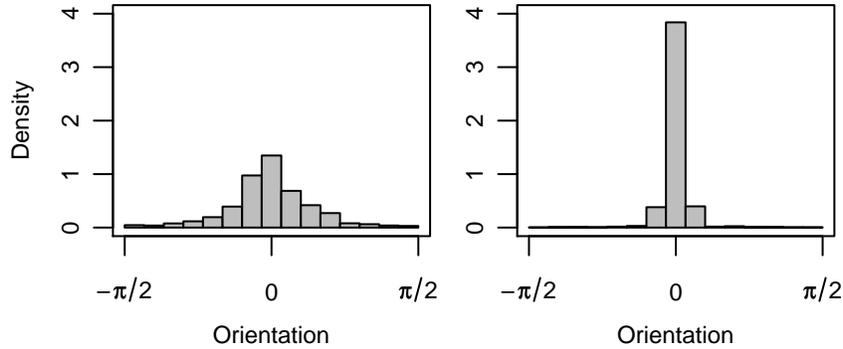
Fig. 4. Histograms of disparity orientation for the raw (left) and epipolar (right) representations. Angles are measured with respect to the horizontal axis of the image. Note that the local epipolar directions are much less variable than the raw disparities.

average vertical disparity increases with distance from the horizontal meridian, causing the local structure to become more variable.

In contrast, the epipolar disparities are quite stable. The smoothing process recovers a structure that resembles a single, average epipolar geometry. In this simulation, the average epipolar lines are parallel, though this is not necessarily always the case. For example, a spatially concentrated distribution of fixation points could produce an asymmetric average map.

It is perhaps surprising that the raw and epipolar disparity maps appear quite similar in figure 5. This can be explained as follows. The difference between the raw and epipolar representations depends largely on the homography that expresses the relative orientation of the eyes. In the present simulation, this homography is not far from the identity, for two reasons. Firstly, we have applied a fixation constraint, which tends to limit the difference in orientation between the views, especially when the scene is relatively distant. Secondly, the field of view over which the homography applies is quite small in this simulation, as described above.

6 Discussion

We have reviewed the measurement of binocular disparity, and shown how it can be represented in relation to the underlying epipolar geometry. The novel contribution of this work is our analysis of the average disparity field, for a fixating observer. We have shown that this contains useful geometric structure, and that this can be extracted by a simple smoothing process.

The most interesting use of the average disparity field is as a *prior* model of the binocular correspondence field. It is straightforward to go from the scatter-matrices $\boldsymbol{D}(\boldsymbol{q})$ defined in section 4 to a probabilistic model of the local disparity
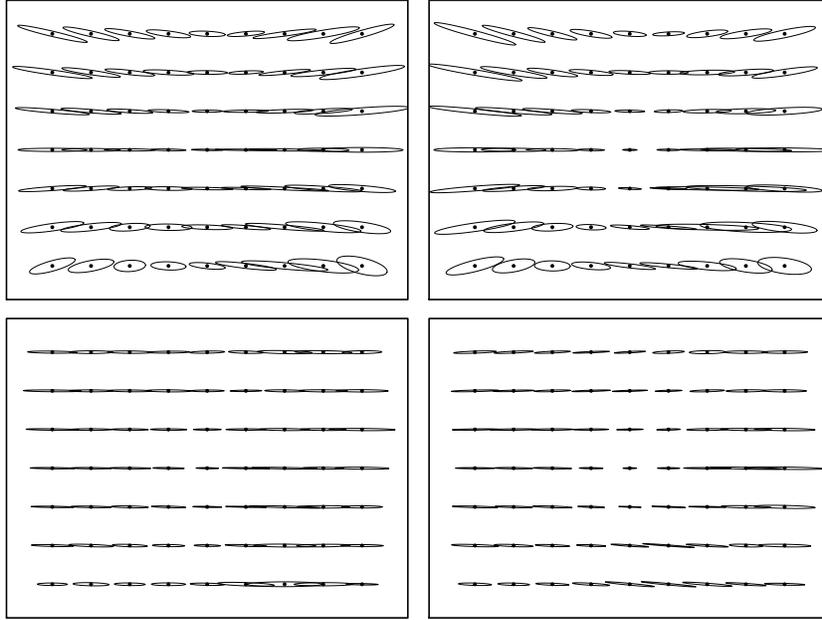
Fig. 5. Structure of the raw (top left & right) and epipolar (bottom left & right) disparity maps, combined over a series of fixations. The maps have been resampled, using the estimator in equation (6), over a region corresponding to the central rectangle that appears in figures 1 & 2. The axes of each ellipse, obtained from the eigen-decomposition of $\boldsymbol{D}(\boldsymbol{q})$, represent the local variability of the disparity field.

vector. This is done via the Mahalanobis distance, which we write as a cost function

$$E\big(\boldsymbol{q}_0, \boldsymbol{q}\big) = \big(\boldsymbol{q} - \boldsymbol{q}_0\big)^\top \boldsymbol{D}(\boldsymbol{q}_0)^{-1}\big(\boldsymbol{q} - \boldsymbol{q}_0\big), \tag{8}$$

where $\boldsymbol{q}$ is the measured feature position, and $\boldsymbol{q}_0$ is the reference point, transferred from the other image, as described in section 1. Hence the candidate disparity is $\boldsymbol{q} - \boldsymbol{q}_0$, with length δ. Recall from section 4 that σ_1^2 and σ_2^2 are the eigenvalues of $\boldsymbol{D}$. It follows that if the disparity is in the characteristic direction, then the cost will be δ/σ_1^2, whereas if it is in the perpendicular direction, the cost will be δ/σ_2^2. The cost is lower in the preferred direction, because $\sigma_1^2 > \sigma_2^2$, assuming that the average disparity has a definite orientation at $\boldsymbol{q}_0$.

These considerations lead directly to a Gaussian model for the prior probability of the match between $\boldsymbol{q}_\ell$ and $\boldsymbol{q}_r$;

$$\mathrm{pr}\big(\boldsymbol{q}_\ell, \boldsymbol{q}_r\big) \propto \exp\!\Big(-\tfrac{1}{2}E_\ell\big(\boldsymbol{H}^{-1}\boldsymbol{q}_r, \boldsymbol{q}_\ell\big) - \tfrac{1}{2}E_r\big(\boldsymbol{H}\boldsymbol{q}_\ell, \boldsymbol{q}_r\big)\Big).$$

The matrix $\boldsymbol{H}$ is the homography that includes the relative orientation of the cameras, as in equation (3). We use both the right-to left and left-to right costs, because the distance defined in equation (8) depends on the average disparity field, and the left and right versions may not be mutually consistent. Here

we have constructed a geometric prior, which depends on the variable orientation of the eyes. This could be readily combined with the photometric prior $\exp\left(-\frac{1}{2}F(\boldsymbol{q}_\ell, \boldsymbol{q}_r)\right)$, which is obtained from the matching-cost F, as defined in equation (4).

In our future work, we plan to incorporate the geometric prior into the image-matching process, as outlined above. We believe that this would improve the estimated binocular correspondences, especially in a biological model based on short-range disparity mechanisms, as described in the introduction. We also plan to evaluate our disparity-smoothing procedure across a wider range of images and fixation points. This will allow us to compare our average correspondence maps to the distribution of disparity-tuned cells in area V1 [1].

Acknowledgments

References

1. Cumming, B. G. 2002: An Unexpected Specialization for Horizontal Disparity in Primate Primary Visual Cortex. *Nature* 418(8), 663–636.
2. Faugeras, O. 1995: Stratification of 3-D Vision: Projective, Affine, and Metric Representations. *Journal of the Optical Society of America A*, 12(3), 465–484.
3. Harris, C. & Stephens, M. 1988: A Combined Corner and Edge Detector. *Proc. 4th Alvey Vision Conference*, 147–151.
4. Huang, J., Lee, A. B. & Mumford, D. 2000: Statistics of Range Images. *Proc. Computer Vision and Pattern Recognition*, 324–331.
5. Knutsson, H. 1989: Representing Local Structure Using Tensors. *Proc. 6th Scandinavian Conference on Image Analysis*, 244–251.
6. Mayhew, J. E. & Longuet-Higgins, H. C. 1982: A Computational Model of Binocular Depth Perception. *Nature* 297, 376–378.
7. Ohzawa, I. & Freeman, R. 1990: Stereoscopic Depth Discrimination in the Visual Cortex: Neurons Ideally Suited as Disparity Detectors. *Science* 249, 1037–1041.
8. Shashua, A. & Navab, N. 1996: Relative Affine Structure: Canonical Model for 3-D from 2-D Geometry and Applications. *IEEE Trans. Pattern Analysis and Machine Intelligence* 18(9), 873–883.
9. Weinshall, D. 1990: Qualitative Depth from Stereo, with Applications. *Computer Vision, Graphics, and Image Processing* 49(2), 222–241.