

An Explicit and Compact Coding of Geometric and Structural Information Applied to Stereo Processing

Norbert Krüger¹, Michael Felsberg², Christian Gebken³ and Martin Pörksen³

¹ Stirling (Scotland), ² Linköping (Sweden), ³ Kiel (Germany)

Abstract

We introduce a compact coding of image information which explicitly separates geometric information (orientation) and structural information (phase and color). We investigate the importance of these factors for stereo matching on a large data set. From these investigation we can conclude that it is *their combination* that gives the best results. Concrete weights for their relative importance are measured.

1 Introduction

In stereo processing with calibrated cameras we can reconstruct 3D points from two 2D points correspondences by computing the point of intersection of the two projective lines generated by the corresponding image points and the optical centers of the cameras (see figure 1 and, e.g., [5]). However, most meaningful image structure is intrinsically one-dimensional [23], i.e., is dominated by edges or lines. Orientation at intrinsically one-dimensional image structures can be estimated robustly and precisely by various methods (see, e.g., [10]). Hence, it is sensible to use orientation information as well for the representation of 3D-information in visual scenes. From two corresponding 2D points with associated orientation (in the following called '2D-line segment'¹) we can reconstruct a 3D point with associated 3D orientation, (see figure 1 and , e.g., [5, 20]): Each of the 2D-lines generated by the points with associated orientation together with the optical center of the camera span a 3D plane (figure 1). Then the intersection of both planes defines the 3D orientation of a 3D-line segment. Only in case that the two planes are identical reconstruction is not possible.

¹Note that we do not want to use *end points* of lines for reconstruction (as, e.g., in [22]). Since they presuppose the *grouping* of points to one connected entity, these tuples are hard to determine non-ambiguously in images. Instead, we make use of the *orientation at a certain image point* which can be estimated with much more precision and robustness.

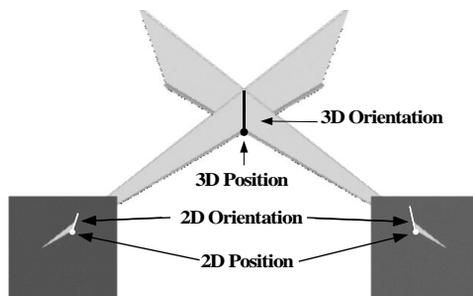


Figure 1: Reconstruction of 3D-point and 3D-orientation from two 2D-point and direction correspondences.

The problem at hand is to find correspondences between image structures in the left and right image. There is severe trouble connected to this problem: Since the scene is seen from different views *the image structure differs* in the left and right image, position of corresponding points are different and the local orientation at these points differ as well (see figure 2). In fact, these differences are the reason why reconstruction is possible: Dissimilarity in position (disparity) determines the depth while dissimilarity in orientation determines the 3D orientation (note that all geometric attributes of an oriented 3D point are covered by its 3D position and 3D orientation). The photometric information going beyond geometric information (in the following called *structural information*) undergoes a complex transformation which has to take into account the transfer of pixel positions (depending on the 3D geometry of the projected object) as well as variation of reflection properties of surfaces according to view variation and the occurrence of occlusion.

For 3D reconstruction we face the following *similarity-dissimilarity dilemma*: We want to find correspondences by *similarity* of the image patches

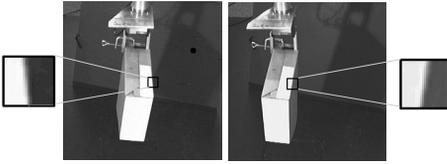


Figure 2: Dissimilarity of corresponding image patches in stereo images.

but we want to reconstruct utilizing their *dissimilarity* in position and the geometric property orientation. While dissimilarity in position can be resolved by a transition of the image patch on the epipolar line the difference in orientation can only be resolved by more complex mechanisms (see, e.g., [7]).

Some stereo similarity functions for intrinsically 1D information use geometric attributes (orientation, length) [2, 19]. However, ambiguity of geometric information leads to a large number of potential matches. Furthermore, significant variation of orientation in both images can occur for entities with small depth (see figure 2). In this paper we will show that we can improve stereo matching significantly by using structural information in addition to geometric information and we give measures for their relative importance.

Alternatively to methods that use geometric information only for feature matching, in classical stereo approaches often a kind of template matching is used to find correspondences in images: Local image patches are compared pixel-wise, see e.g., [21, 14]. These methods are also called ‘area-based stereo’ or ‘intensity based correspondence analysis’ [12]. The similarity function might be a (normalized) squared error (see [21]) or a (normalized) cross correlation (see [12]). In these approaches the above mentioned similarity–dissimilarity dilemma is not treated explicitly. Indeed, the underlying argument or ‘hope’ is that despite the deviation of orientation the template match is sufficiently close, a hope which is not necessarily justified (see figure 2). This ‘hope’ is with high likelihood fulfilled, when the depth in relation to the basis width of the stereo rig is sufficiently large. The performance of matching procedures will decline, when orientation difference is too large since they do not distinguish the two above-mentioned factors. Furthermore, they are not able to make use of 3D–orientation information since they do not represent 2D–orientations explicitly.

Some authors use both factors, orientation and

structural information. In [7] variation of the local image patches are taken into account explicitly by applying an affine transformations of the grey values of the image patch. The parameters of this affine transformation have to be computed by finding a solution of an overdetermined set of equations. Once these parameters are known, relative orientation difference of the image patches can be used for reconstruction. Of course, solving the set of equations can be a time demanding procedure. Taking assumptions about the 3D geometry into account (more specifically, assuming the edge being produced by the intersection of planes) the complexity of the affine transformation can be reduced [20] but still an optimization method has to be applied. Other problems concerned with this approach are that the assumption of plane surfaces is not necessarily full-filled. Furthermore, for edges caused by intersection of strictly homogenous 3D–surfaces an optimal transformation can not be computed. Finally and most importantly, from the point of view of object representation *a more compact storage of structural information than the image patch itself is wanted.*

In this paper, we introduce a similarity function which uses geometric information (orientation) and structural information in a direct way, i.e., without the need of solving a set of equations. An intrinsically one-dimensional structure in a grey level image can be described by orientation (or geometric information) and information about its structure (e.g., it can be distinguished between being a dark/bright (bright/dark) edge or a bright (dark) line on dark (bright) background). Of course, there is a continuum between these different grey level structure. The local phase as additional feature allows to take the grey level information into account (as one parameter in addition to orientation) in a very compact way (see, e.g., [9, 15, 6]).

As it was shown by e.g., [14, 11] color also is an important cue to improve stereo matching. We use color triplets to describe the left and right side of the edge in RGB space: Color at the left and right side of an edge is averaged to two color vectors indicating the mean color structure of two half sides (see figure 3).²

The paper is structured as following: In section 2 we briefly describe our feature processing. In image processing we are able to compute a local *ori-*

²More precisely we use for signal patches corresponding to lines (i.e., phase close to 0 or π) also a color value for the center line. For the sake of simplicity we neglect this in the following description.

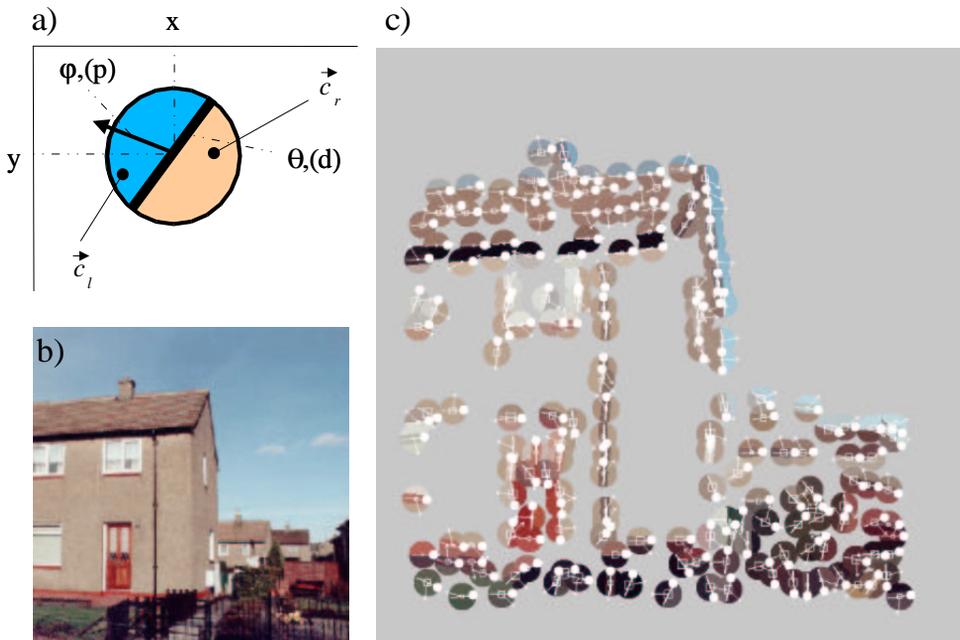


Figure 3: **Top left:** Schematic representation of a basic feature vector. Position is coded by (x, y) , orientation by θ (or direction as d respectively), phase by φ (or p when associated with a direction), and color by (\vec{c}_l, \vec{c}_r) . **Bottom left:** Frame in an image. **Right:** Extracted feature vectors.

entation. However, taking more global interdependencies (such as consistency across different views) into account we can extend the concept of orientation to the concept of *direction*. In section 3 we therefore discuss the concept of direction in more detail. A similarity function is derived in section 4 that allows to steer explicitly the influence of the orientation deviation in contrast to the structural information and also the influence of phase versus color information. The relative importance of orientation, phase and color is investigated in section 5.

We would like to point out that it is not our aim to derive a perfect stereo system. Stereo is an ambiguous visual modality since the correspondence problem can become extremely awkward in complex scenes and mismatches lead to wrong 3D estimates. Integration of other visual modalities (see, e.g., [1, 18, 3]) and integration of ambiguous information over time (see, e.g. [4, 13, 20, 16]) has to be used to achieve robust information. However, the aim of this paper is to define and investigate an appropriate local similarity function which makes use of structural and geometric information and to derive statements about the relative importance of

geometric versus structural information, and phase versus color information.

2 Feature Processing

In this section we describe the processing of information (orientation, phase and color) used in our stereo algorithm. Note that in [18] the same kind of features are used to determine their statistical relationship.

Position, Orientation and Phase: In this paper we will use a systematic mathematical description of geometric and structural information of grey level images based on the monogenic signal [6]. The monogenic signal performs a *split of identity*, i.e., it orthogonally divides the signal into energetic information (indicating the likelihood of the presence of a structure), its orientation θ and its structure (expressed in the phase φ). Features are extracted at energy maxima in local image patches where the position is parameterized by \vec{x} (see figure 3). The variance of orientation in an image patch (computed from pixel positions of high energy) is indicated as a square in the displays of feature vectors in figure 3 (right). In our simulations we only use features

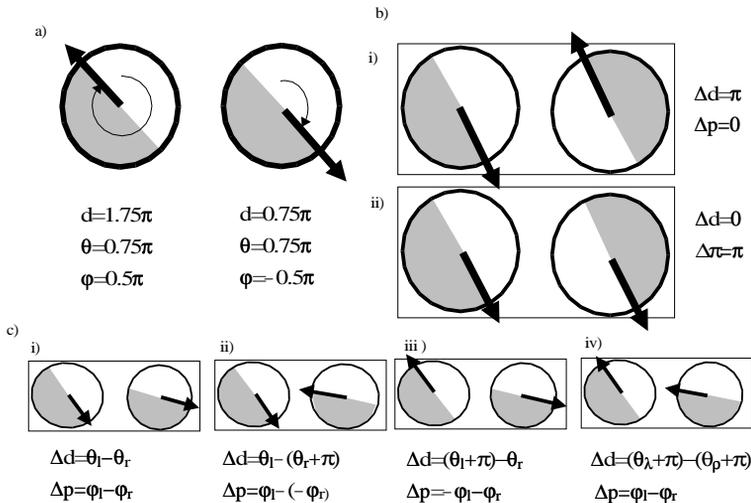


Figure 4: a) Given extracted orientation θ and phase φ the same image patch can be interpreted in terms of direction d and phase p as ($d = \theta + \pi, p = -\varphi$) (left) or ($d = \theta, p = \varphi$) (right). b) The similarity in structure and orientation of two image patches changes with the interpretation of direction. c) Theoretically possible interpretations of direction in the left and right image patch. Note that ii) and iii) are geometrically not possible according to the Stereo coherence constraint.

for which the variance of orientation within a small patch is below and the magnitude is above certain thresholds, i.e., features that correspond to image patches of intrinsic dimension close to one.

The phase can be used to interpret the kind of contrast transition at this maximum [15], e.g., a phase of $\frac{\pi}{2}$ corresponds to a dark-bright edge, while a phase of 0 corresponds to a bright line on dark background. The continuum of contrast transition at an intrinsic one-dimensional signal patch can be expressed by the continuum of phases.

Color: The distribution of phases in natural images has been investigated in [18]. There exist clear peaks at $\varphi = \pi/2$ and $\varphi = -\pi/2$ which show that edges (i.e., intrinsic 1-dimensional signals with odd symmetry) are the dominant one-dimensional structure in natural images while line structures (i.e., intrinsic 1-dimensional signals with even symmetry) are less dominant. Our model for an intrinsically one-dimensional signal patch (see figure 3) therefore describes edges.³

³Although there is significantly more edge like structures than line like structures in natural images we can also make use of an extra line model to describe intrinsically one-dimensional image patches with phase close to 0 or π . The introduction of this model makes only small difference for stereo matching (but is important in other contexts). We neglect this issue here.

To integrate the modality color at intrinsically one-dimensional image structures we perform an averaging in the RGB color space over the left and right part ('left' and 'right' defined by the associated line segment) of the image patch (see figure 3)⁴.

We get two vectors $\vec{c}_l = (c_r^l, c_g^l, c_b^l)$ and $\vec{c}_r = (c_r^r, c_g^r, c_b^r)$, representing the red, green and blue values of the left and right side of the edge.

Therefore, our basic feature vector has the form

$$\vec{f} = (\vec{x}, \theta, \varphi, (\vec{c}_l, \vec{c}_r)).$$

3 Direction and Orientation

In the feature processing described in section 2 we extract orientation information θ which takes values in $[0, \pi)$. However, when we add structural information at the local edge (see figure 4a) we can extend the concept of orientation to the concept of direction (see figure 4a and [9]), parameterized by $d \in [0, 2\pi)$ (Note that the local phase can change as well when we go from orientation to direction. We denote this corrected phase p instead of φ). Although for each single edge in figure 4a) or 4c) two interpretations for direction are possible we can

⁴The image patch has a radius of 8 pixels.

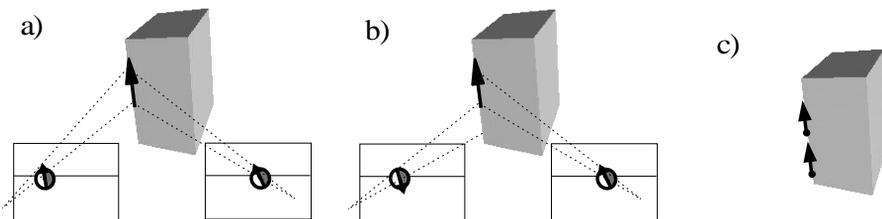


Figure 5: a) Geometrically possible interpretation of direction. b) Geometrically impossible interpretation of direction. c) Good Continuation.

overcome this ambiguity by taking global relations into account. For example, by assuming both edges as part of the same 3D Gestalt (see figure 5c) both assignments of direction have to be coherent.

In the stereo domain, the association of direction to two corresponding 2D line segments implies a 3D direction (see figure 5a). On the other hand, a 3D direction implies 2D directions of its projections (see figure 5a). We call this relation the *direction uniqueness constraint*.

Another constraint is *Stereo Direction Coherence*: Assuming parallel cameras. If for a line segment l^l in the left image holds $d < 1/2\pi$ or $d > 3/2\pi$ than for the corresponding line segment in the right image must hold the same. Figure 5a shows a valid interpretation of direction while figure 5b shows a non-valid interpretation of direction.⁵

Both constraints are used in the stereo similarity function described in section 4. The stereo direction constraint is used to reduce the number of possible correspondences of directed line segments. The direction uniqueness constraint states that reconstruction of directed 3D line segments is possible. It is one advantage of our similarity function that it associates a direction to line segments, i.e., can be used to overcome the ambiguity of local direction estimation.

4 A new Stereo Similarity Function

Our basic local features at position \vec{x} can now be extended to $\vec{f} = (\vec{x}, d, p, (\vec{c}_l, \vec{c}_r))$, $d \in [0, 2\pi]$ representing the direction, $p \in [-\pi, \pi]$ representing the phase ($p = \varphi$ if $d = \theta$ and $p = -\varphi$ if $d = \theta + \pi$),

⁵In case of non-parallel cameras the constraint reads: If for a line segment l^l in the left image holds that the vector of direction points above the epipolar line in the left image (constituted by the corresponding line segments in both images) than for the vector of direction of the corresponding line segment in the right image holds the same and vice versa.

(\vec{c}_l, \vec{c}_r) , $\vec{c}_i \in [0, 1] \times [0, 1] \times [0, 1]$ represent the color in RGB space. Since we want to neglect information about the magnitude we ensure $\|\vec{c}_i\| = 1$.

A straightforward distance function between two line segments for stereo matching is the weighted sum of differences of the orientation and the structural information associated to the features extracted from the left and right image:

$$\mathcal{D}(\vec{f}^l, \vec{f}^r) = \quad (1)$$

$$\alpha(\Delta d) + (1 - \alpha)(\beta(\Delta p) + (1 - \beta)(\Delta c)).$$

$\alpha \in [0, 1]$ represents the weight for the geometric information compared to structural information, $\beta \in [0, 1]$ represents the weight for phase compared to color information. Δd , Δp and Δc are all defined such that they take value in $[0, 1]$. Here we want to remark that all values Δd , Δp , Δc are normalized such that they have comparable mean and standard deviation according to [8].

The concept of direction is essential for the definition of structural information, since the structural part switches under a rotation of π (see figure 4a) and the two triplets (\vec{c}_l, \vec{c}_r) switch as well). To define a stereo similarity function the concept of direction is essential as well. For instance, the image patches in figure 4b(i) (if direction is interpreted as indicated) have same structure but opposite direction while the image patches in figure 4b(ii) have same direction but different structure.

However, in images only the orientation is locally measurable only while the structural part switches under the assumed underlying direction (see figure 4b and [9]). Although from a more global perspective consistent interpretation of direction could be associated to the line segments (see section 3) we can not decide locally which interpretation of direction is appropriate. Therefore, to compare two feature vectors $(d^l, p^l, (\vec{c}_l^l, \vec{c}_r^l))$, $(d^r, p^r, (\vec{c}_l^r, \vec{c}_r^r))$ in the stereo similarity function we have to look at

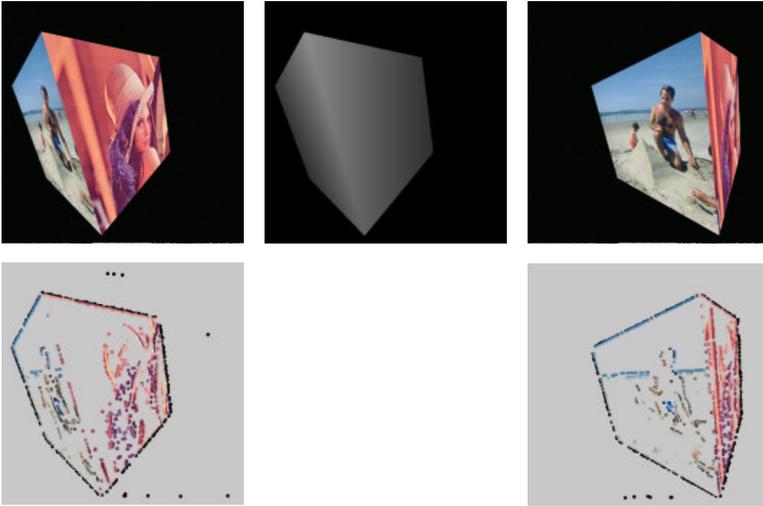


Figure 6: Top: Left Image (left), ground truth (middle), right image (right). Bottom: Extracted features in the left and right image.

all geometrically possible interpretations of direction for the left and right image patch. We have to deal with four cases:

- 1) The measured orientation in the left and right image equals the underlying direction: $d^l = \theta^l$, $d^r = \theta^r$ (see figure 4ci).
- 2) The measured orientation in the left image equals the underlying direction but the measured orientation in the right image is related to the direction by $d^r = \theta^r + \pi$. This also implies that the phase has opposite sign than the locally measured phase ($p^r = -\varphi^r$) and the color triplets switch as well (see figure 4cii).
- 3) The measured orientation in the left image equals is related to the direction by ($d^l = \theta^l + \pi$) which also implies that the phase has opposite sign than the locally measured phase $p^l = -\varphi^l$ (see figure 4ciii) and the color triplets switch as well. The underlying direction in the right image equals the measured orientation.
- 4) The measured orientation in the left image is associated to the direction $d^l = \theta^l + \pi$ and the measured orientation in the right image is associated to the direction $d^r = \theta^r + \pi$.

The exact definition of the stereo similarity function can be derived by treating and combining these four cases in an appropriate way (for details, see [17]).

5 Experiments

In this section we investigate the relative importance of geometrical versus structural information as well as the relative importance of phase versus color. That means we investigate the quality of stereo matching depending on the weights α and β in (1).

Measuring performance: To achieve statistically relevant statements about the importance of the different factors of visual information we need a large data base. Since a manual generation of ground truth from natural images is extremely tedious we use images created in a virtual environment by texture mapping with natural images: Natural images are mapped onto a randomly rotated cube (see figure 6). This ensures that we deal with data close to natural conditions as well as a known 3D-structure of the scene.

Our measure of performance is the number of 3D-line segments with associated disparity close to the ground truth (in our case we chose a deviation of 3 pixel, $t = 3$) divided by the number of extracted 3D-line segments.

In many stereo algorithms additional constraints are used to improve performance. The *epipolar line constraint* says that the corresponding point to a point in the left image must be on a epipolar line in the right image. This constraint is always valid (see, e.g., [5, 12]). The *uniqueness constraint* states

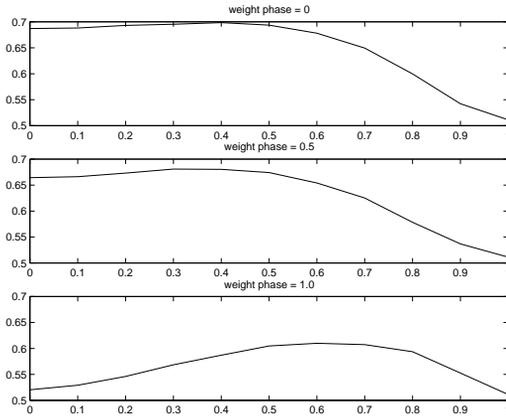


Figure 7: Slices of surface shown in figure 8. The x-axis represent α . Top: $\beta = 0$. Middle: $\beta = 0.5$. Bottom: $\beta = 1.0$.

that a 3D-point can not have two distinct projections in an image. This constraint is always valid as well. Other constraints such as *ordering* (i.e., for a point which is left to another point the corresponding points in the right image have to have the same order) and *restrictions on the absolute disparity* are only valid in most circumstances but there exist geometric exceptions.

In our simulations we only make use of the epipolar constraints and the uniqueness constraint but we do not use any kind of further restrictions to improve stereo since our aim is not come up with a system which optimal performance but to investigate the different factors of visual information according to their contribution to stereo matching. We use a set of 40 images. The chance level (i.e., when our similarities are defined randomly) of performance is 30.6%. The performance with a normalized cross correlation⁶ on the grey level image is 67.7%, the performance with a normalized cross correlation (by adding the results of correlation in each sub-channel) on the color level image is 68.5% .

Contribution of structural versus geometric information in grey level images: Figure 7 (bottom) shows the variation of performance on the test set of 40 images for different α when we set $\beta = 1.0$, i.e., we use only phase and no color. We recognize a peak performance of 61% for $\alpha = 0.6$. We see

⁶The comparisons are made at the very same pixel positions for normalized cross correlation than for our new similarity function.

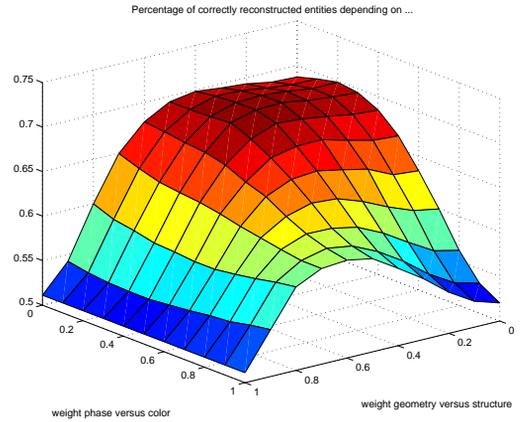


Figure 8: Performance for varying α and β . α represents the 'weight geometry versus structure' and β the 'weight phase versus color'.

that optimal performance is achieved by *combining* structural and geometric information (using direction only we get only 51.1% and using phase only we get 52%). Here, the importance of geometric information for stereo matching is slightly higher than for the structural information coded in the phase. The performance is lower than for a normalized-cross correlation matching with 10×10 patches. However, we achieve a reasonable performance although we *reduce the signal to 2 parameters only (direction and phase) compared to 100 parameters for the normalized cross correlation!* This corresponds to a reduction by a factor of 50.

Contribution of structural versus geometric information using also color information: Figure 7 (top) shows the variation of performance performance on the test set different α when we set $\beta = 0.0$, i.e., we use color only as structural information. We see again that optimal performance is achieved by combining geometrical and structural information. We also recognize that the structural information coded in color leads to better results than by using phase only.

Figure 8 shows the performance when we vary α and β . The plots in figure 7 are slices of this figure. We achieve a top performance of 70.5% for $\alpha = 0.3, \beta = 0.3$. Once again we see that the *combination* of structural and geometric information gives optimal performance and we can also recognize that the *combination* of phase and color information gives the best results, i.e., both factors can be used complementary. Our parameteriza-

tion of color leads to an increase of performance by 9.5% compared to the use of grey level information only. Structural information can now be weighted higher compared to grey level information (0.7 versus 0.4). However, since our similarity function distinguishes between left and right side of the edge some geometric information is coded as well in the color triplets.

For the optimal weights we achieve on our data set a even higher performance (70.5% versus 68.5%) than using normalized cross correlation for color images although we reduce a 100 dimensional image patch to 8 parameters only.

6 Conclusion

We have investigated a compact and explicit coding of local image information for intrinsically one-dimensional signals in terms of direction, phase and color attributes. We applied applying this coding within the stereo domain. By making use of the explicit separation of geometric and structural information we could compute the relative importance of the different sub-aspects for stereo processing. We could show that it is the combination of aspects that gives the best results. Moreover, we could show that we can achieve high matching performance although we reduce the image information by a factor of 50 for grey level images. For color images we can achieve an even higher matching performance than with a normalized cross correlation although we reduce the image information by a factor of more than 35. Therefore, this compact coding also promises efficient applicability for tasks that require low storage costs, such as, e.g., object coding and object recognition.

Acknowledgment: We would like to thank Wolfgang Förstner, Georgy Gimelfarb, Reinhard Koch, Gerald Sommer and Andrew Zisserman for fruitful discussions.

References

[1] J. Aloimonos and D. Shulman. *Integration of Visual Modules — An extension of the Marr Paradigm*. Academic Press, London, 1989.

[2] N. Ayache. *Stereovision and Sensor Fusion*. MIT Press, 1990.

[3] A. Cozzi and F. Wörgötter. Comvis: A communication framework for computer vision. *IJCV*, 41:183–194, 2001.

[4] O. Faugeras and L. Robert. What can two images tell us about the third one? *International Journal of Computer Vision*, 18(1), 1996.

[5] O.D. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, 1993.

[6] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 41(12), 2001.

[7] W. Förstner. Image matching. In R.M. Haralick and L.G. Shapiro, editors. *Computer and Robot Vision*. Addison Wesley, 1993.

[8] K. Fukunaga, editor. *Introduction to statistical pattern recognition (2nd ed)*. Academic Press, 1990.

[9] G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, Dordrecht, 1995.

[10] B. Jähne. *Digital Image Processing – Concepts, Algorithms, and Scientific Applications*. Springer, 1997.

[11] J.R. Jordan and A.C. Bovik. Using chromatic information in edge based stereo correspondence. *Computer Vision, Graphics and Image Processing: Image Understanding*, 54:98–118, 1991.

[12] R. Klette, K. Schlüns, and A. Koschan. *Computer Vision - Three-Dimensional Data from Images*. Springer, 1998.

[13] R. Koch. Model-based 3-d scene analysis from stereoscopic image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 49(5), 1994.

[14] A. Koschan. How to utilize color information in dense stereo matching and in edge based stereo matching? *Proceedings of ICARCV*, 1994.

[15] P. Kovési. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3), 1999.

[16] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 13(2), 2002.

[17] N. Krüger and M. Felsberg. A new stereo algorithm that integrates geometrical and structural information. *to be submitted*.

[18] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *Submitted*.

[19] G. Medioni and R. Nevatia. Segment-based stereo matching. *Computer Vision, Graphics and Image Processing*, 31, 1985.

[20] C. Schmid and A. Zisserman. Automatic line matching across views. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[21] Y. Shiray. *Three-dimensional Computer Vision*. Springer (Berlin), 1987.

[22] R.K.K. Yip and W.P. Ho. A multi-level dynamic programming method for stereo line matching. *Pattern Recognition Letters*, 19, 1998.

[23] C. Zetsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30, 1990.