Multi-Modal Matching Applied to Stereo

Nicolas Pugeault[1], Norbert Krüger[1,2]
(nicolas@cn.stir.ac.uk, norbert@cn.stir.ac.uk)
[1] University of Stirling, Department of Psychology
Computational Neuroscience - Computer Vision Group
Scotland, UK

[2] Aalborg University, Esbjerg
Denmark

Abstract

We introduce a compact coding of image information which explicitly separates visual information into geometric information (orientation) and structural information (phase and colour) and temporal information (optic flow). We investigate the importance of these visual attributes for stereo matching on a large data set. From these investigation we can conclude that it is *the combination* of different attributes that gives the best results. Concrete weights for the relative importance of different visual attributes are statistically determined.

1 Introduction

In stereo processing with calibrated cameras we can reconstruct a 3D point from two 2D point correspondences or from two corresponding 2D points with associated orientation, we can reconstruct a 3D point with associated 3D orientation, (e.g. [6, 24]). The problem at hand is to find correspondences between image structures in the left and right image.

To find correspondences, stereo similarity functions between image patches or features in the left and right image need to be defined. Some similarity functions use geometric attributes (such as, orientation or length) [2, 21]. 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. Alternatively to methods that use geometric information only for feature matching, some authors use both factors, orientation and structural information. For example, in [8] variations of the local image patches are taken into account explicitly by applying an affine transformation of the image patch grey values. The parameters of this affine transformation have to be computed by finding a solution of an over-determined 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. Making 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 [24] 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

BMVC 2003 doi:10.5244/C.17.28

full-filled. Furthermore, for edges caused by intersection of strictly homogeneous 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 that makes use of geometric and structural information in a direct way, *i.e.* without the need of solving a set of equations. To improve stereo matching we also use colour and temporal information. In [10, 12] it has been shown that the use of colour can improve stereo matching significantly. Our work confirms this result. Going beyond [10, 12], we are able to give a statement about the relative importance of colour compared to other visual modalities. We make further use of temporal information in terms of the optic flow.

Our similarity function is based on multi–modal image descriptors (see figure 1 and [19]) that covers geometric information (orientation), structural information (phase), colour and temporal information (optic flow). We will show that the use of multiple modalities improves stereo matching performance. Since our similarity function explicitly steers the influence of the different visual modalities, we are able to give concrete weights for their relative importance. We can also show that optimal weights are reasonably robust over different scenes.

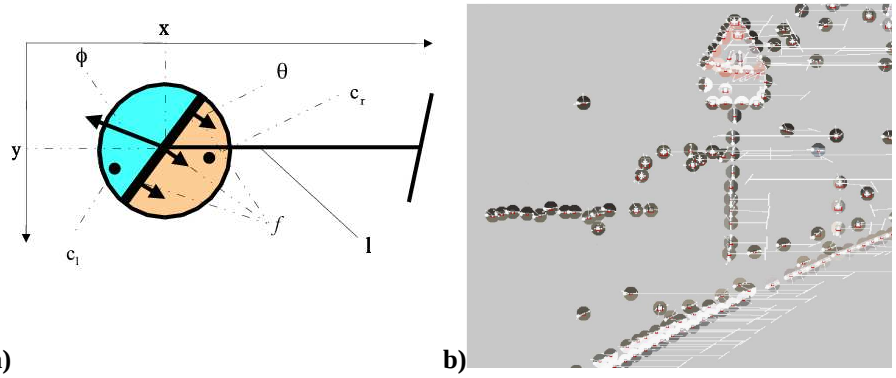
We would like to point out that it is not our aim to derive a perfect stereo system. Stereo is an ambiguous visual modality since the correspondence problem can become extremely awkward in complex scenes and mismatches lead to wrong 3D estimates. Integration of other visual modalities (see, e.g., [1, 20, 4]) and integration over time (see, e.g. [5, 11, 24, 15]) 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 multiple aspects in visual scenes. We derive statements about the relative importance of the different visual aspects. Finally and most importantly, we show (by comparison to a normalized cross–correlation comparison) that our image representation leads to a condensation of information (up to a factor of 96.6%) while preserving the relevant information.

The paper is structured as following: In section 2, we briefly describe our feature processing. A distance function for optic flow vectors is described in section 3. Using this, we integrate the optic flow in a similarity function that also covers orientation, phase, and colour. This similarity functions allows us to steer explicitly the influence of the different visual attributes. The relative importance of orientation, phase, colour and optic flow is investigated in section 4.

2 Feature Processing

In this section we describe the processing of information (orientation, phase, colour and optic flow) used in our stereo algorithm. Note that in [20] the same kind of features are used to determine their statistical relationship in natural images.

We will use a systematic mathematical description of geometric and structural information of grey level images based on the monogenic signal [7]. 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 in local image patches which position is parameterized by $X = (x,y)$ (see figure 1**a)**). In our simulations we only use



a) **b)**

Figure 1: **a)** Schematic representation of a basic feature vector. Position is coded by (x,y), orientation by θ, phase by ϕ, and colour by (c_l,c_r), the colour on both sides of the edge. l is the disparity between the primitive and its match in the other image. . **b)** Here the previously described primitives are extracted from an image. The white lines represent the disparities l for all the primitives and point to the position of the matching primitive in the other image.

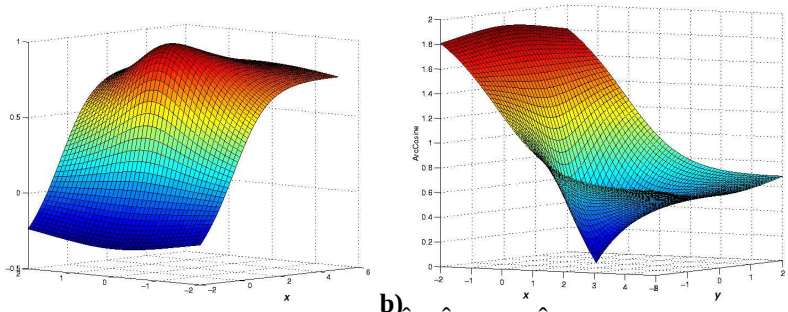
features for which the variance of orientation within a small patch is below and the magnitude is above certain thresholds, *i.e.* features that correspond to image patches of intrinsic dimension close to one, since orientation and phase are only defined for intrinsicly one-dimensional signals (*c.f.* [16]). The phase ϕ can be used to interpret the kind of contrast transition at this maximum [14], e.g., a phase of $\frac{\pi}{2}$ corresponds to a dark–bright edge, while a phase of zero 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. The local phase as additional feature allows us to code structural grey level information into account (as one parameter in addition to orientation) in a very compact way (see, e.g., [9, 14, 7]).

As it was shown by e.g. [10, 13], colour is also an important cue to improve stereo matching. The pixel data of the image contains the three components red, green and blue. As we are already using the intensity information through the phase, we want a colour vector excluding this information. We decide to use the YUV colour space (cf. [23]), Y containing the intensity information, and U and V coding the colour. This allows us to reduce the colour information from 3 to 2 dimensions with a simple linear transformation.

The colour information of a primitive is defined by the colour on both sides of an edge, and, in the case of a line structure (if $\phi \simeq 0$ or $\phi \simeq \pi$), the colour of the line itself. The colour information vector is then $C = (c_l, c_m, c_r)$. The three component vectors $c_l = (c_U^l, c_V^l)$, $c_m = (c_U^m, c_V^m)$ and $c_r = (c_U^r, c_V^r)$ with $c_j^i \in [0,1]$ hold the U and V values of the left side, the center and the right side of the edge. Consequently the colour information we are using is *6–dimensional*.

To this feature description we add the optic flow local measurement, using the well known *Nagel* algorithm (cf. [22]).

As a result we got a multimodal visual primitive that gives a rich but condensed description of a local image patch. For more details concerning this kind of image representation we refer to [18].



a) **b)**

Figure 2: **a)** Graph of the dot product $\hat{f}_1 \cdot \hat{f}_2$, with $\hat{f}_1$ being the normalized 3D vector equivalent to the 2D vector $f_1 = (x,y)$, $x \in [-2,+5]$ and $y \in [-2,+2]$, and $\hat{f}_2$ being the normalized 3D vector equivalent to the 2D vector $f_2 = (1,0)$: $\hat{f}_2 \simeq (0.7071, 0, 0.7071)$. **b)** Graph of the distance function $d(f_1, f_2)$.

The resulting primitives are represented by the following vector:

$$E = (X, \theta, \phi, (c_l, c_m, c_r), f) \tag{1}$$

With $X = (x,y)$ being the position of the primitive in the image, $\theta \in [0, 2\pi]$ the orientation and $\phi \in [-\pi, \pi]$ the phase. Finally $f = (u,v)$ is the optic flow vector at this location.

3 A Multi-Modal Similarity Function

To address the problem of stereo correspondances, we need to define a metric to estimate the quality of a match between two local primitives E and E' (being primitives as defined in equation (1)). A similarity function involving measures of the distances in orientation $d_\theta(E,E')$, phase $d_\phi(E,E')$ and colour $d_c(E,E')$ of the primitives has already been proposed in [17]. Here we extend this similarity function, including our optic flow distance $d_f(E,E')$.

For the optic flow information to be integrated in the stereo correspondances discrimination, a distance metric between any pair of optic flow vectors (f_1, f_2) has to be defined. The vectors may be dissimilar in length or orientation. We want a similarity function so that the vectors have a low similarity if their orientation is widely different. If the orientation is close, then the vector would have a higher similarity if their lengths are close. The dot product of the normalized two vectors is proposed as distance for vectors by [3]. If for a vector $f = (x,y)$ we consider the equivalent homogenous 3D vector $f_{3D} = (x,y,1)$, then the normalized homogeneous vector is: $\hat{f} = (\hat{x}, \hat{y}, \hat{z}) = \frac{f_{3D}}{||f_{3D}||}$, so that $||\hat{f}|| = 1$. Then the dot product of the normalized 3D equivalent of two vectors gives a possible value for those two vectors similarity:

$$sim(f_1, f_2) = \hat{f}_1 \cdot \hat{f}_2 \tag{2}$$

This formula allows comparison for length as well as orientation: high difference in orientation (more than 45 degrees) yields a very low similarity whatever the length of the vectors, which is consistent with our perception of optic flow similarity. The use of normalized 3D vectors assure a consistent behaviour while comparing vectors of any size range.

The graph 2**a)** shows this function 2 for vectors of coordinates $f_1 = (x,y)$ with the vector $f_2 = (1,0)$. This curve is effectively a representation of the similarity of two vectors. Similar vectors have a high value (up to one for identity), also vectors sharing a close

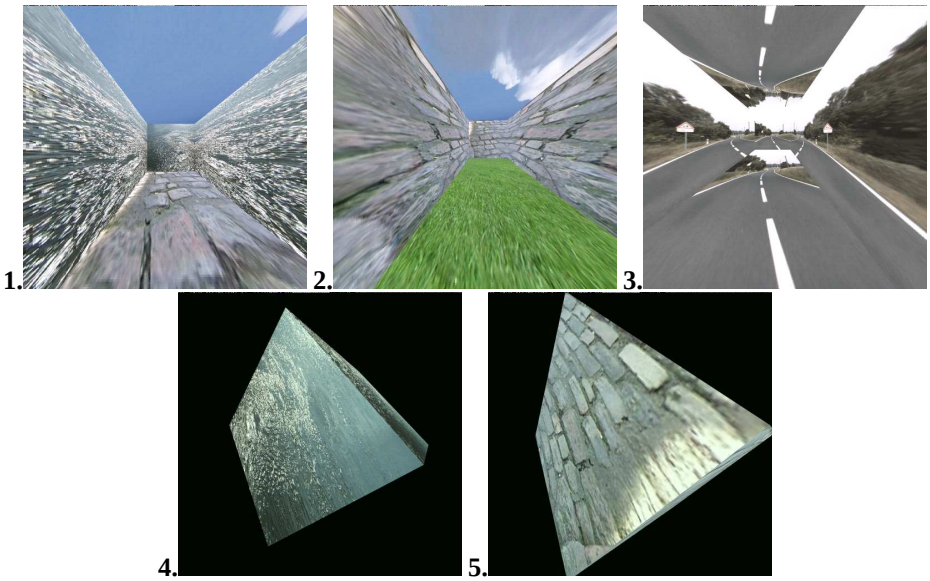


Figure 3: The five scenes used for the test.

orientation keep a higher similarity while the function value reduces sharply for vectors of widely divergent orientation

In order to get a distance funtion between the optic flow vector, and to improve the steepness of the curve close to the identity we apply the ArcCosine function to 2. Our distance becomes:

$$d(f_1, f_2) = ArcCos(\hat{f}_1 \cdot \hat{f}_2) \tag{3}$$

The high steepness of this function (cf. figure 2**b)**) allows us to identify the best match in a set of closely related vectors.

The resulting similarity function can be written as follows:

$$D_w(E,E') = w_\theta d_\theta(E,E') + w_\phi d_\phi(E,E') + w_c d_c(E,E') + w_f d_f(E,E') \tag{4}$$

with $w = (w_\theta, w_\phi, w_c, w_f)$ the weighting of the modalities distances between the two primitives so that $w_\theta, w_\phi, w_c, w_f \in [0,1]$ and $w_\theta + w_\phi + w_c + w_f = 1$

All the modalities measured for those local primitives have very different nature and distribution. As we want to combine them we need to normalize them somehow beforehand. We applied a normalization function proposed in [25].

4 Results

In this section, we investigate the relative importance of the modalities defining a primitive (as in equation 1) for the task of stereo correspondances identification.

Concretely it means the quality of the stereo matching obtained using the similarity function defined in section 3, depending on the weights of each modality.

4.1 Data

We tested the quality of our stereo matching correspondances using artificial 3D scenes with natural textures (figure 3). The scenes feature a camera motion along a textured

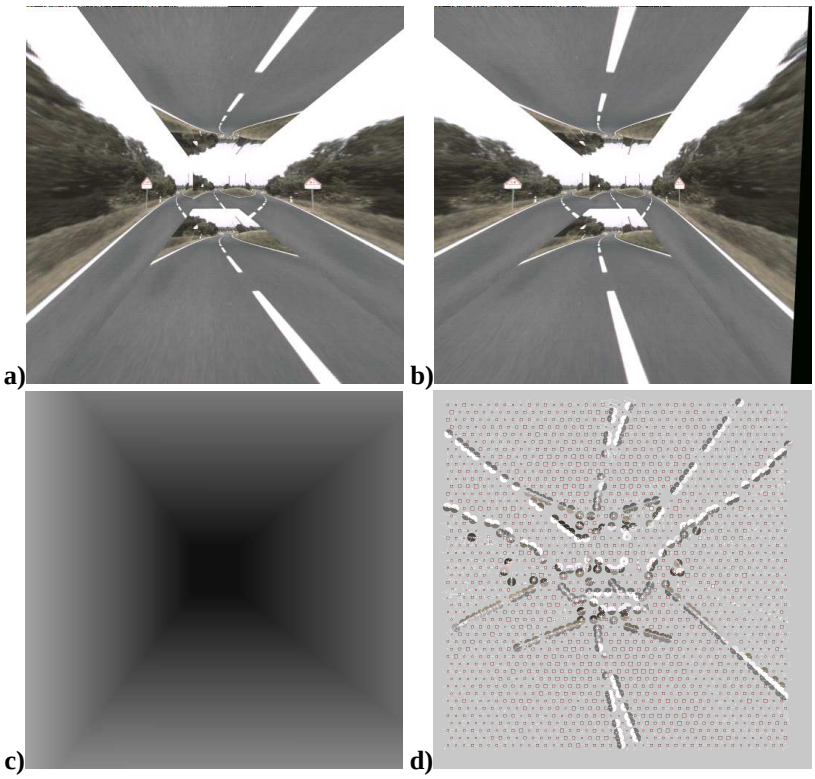


Figure 4: **a)** and **b)** are respectively the left and right images of the scene. **c)** shows a greymap of the disparity ground truth for this frame. **d)** shows the features extracted from this frame.

corridor, or rotating cubes, with varying textures.

On one hand, those scenes provide us an accurate ground truth for the scene depth (knowing the exact scene layout, camera projection matrices and motion), and so an exact measure of the theoretical disparity can be computed. On the other hand, the projected textures ensure that we do work with natural structures.

By comparing the estimation of the disparity found with our method we can have a measure of the performance of the similarity function for this task. We consider sequences of 10 frames for each of those sequences, which comes to a total of 50 stereo frames of 512 per 512 pixels. Our statistics are made over a total of more than 66,000 matches.

In order to compare the relative importance of those modalities, we define a relative weighting $\alpha, \beta, \gamma \in [0,1]$. α is the relative weight of the optic flow versus all the static modalities, β the weight of geometric information (the orientation measurement) versus structural information (phase and colour) and finally γ is the relative weight of phase versus colour.

We reformulate the distance (4) to use those relative parameters:

$$D'_{\alpha,\beta,\gamma}(E,E') = \alpha d_f(E,E') + (1-\alpha)(\beta d_\theta(E,E') + (1-\beta)(\gamma d_\phi(E,E') + 1 - \gamma d_c(E,E'))) \quad (5)$$

from (5) we define the similarity as follows:

$$Sim_{\alpha,\beta,\gamma}(E,E') = 1 - D'_{\alpha,\beta,\gamma}(E,E') \quad (6)$$

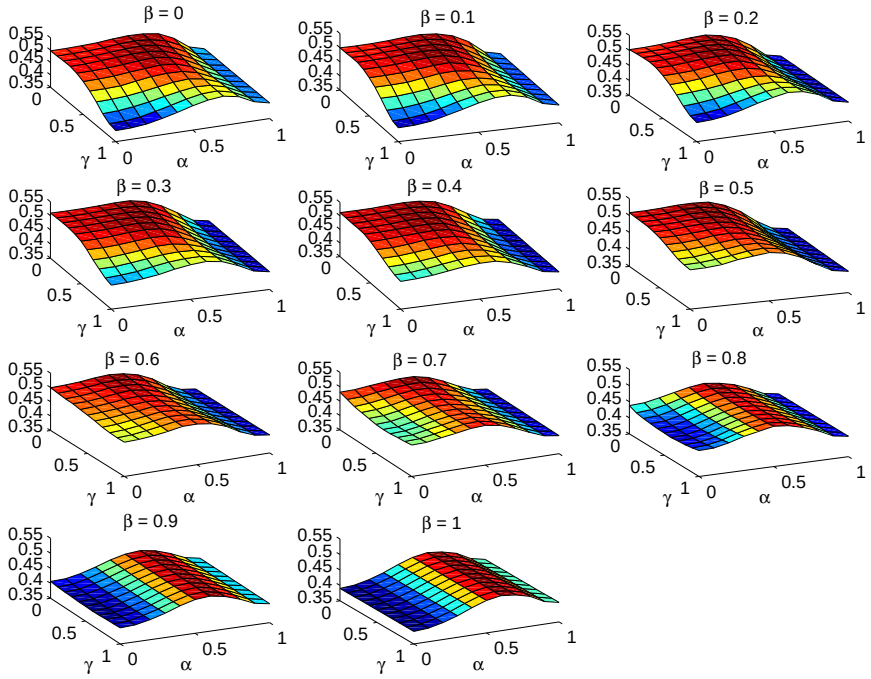


Figure 5: Graph of the disparity quality for different modality weights over all sequences. The different graphs are for different values of β, the α values are along the x axis and the γ along the y.

This formula is used to identify the best corresponding local primitive of the right image (maximizing (6)) along the epipolar line (cf. [6]). The subsequent disparity is then compared to the ground truth for the disparity of the sequence. The quality of the similarity function is then evaluated simply by the ratio of correct correspondances over all matches.

4.2 Performances using all Modalities

To have a performance baseline to estimate the quality of our correspondances, we calculated the chance performance (the performance using a random similarity function) and a cross-correlation over 10x10 patches (here the similarity function used is the cross correlation of the patches). Those have been calculated for our five benchmark sequences.

The matching performance of our similarity function for values of $\alpha, \beta, \gamma \in [0,1]$ is shown in figure 5. We can see a plateau for α close to 1. In this case, only the optic flow modality is being used, so the variations in β and γ do not affect the surface. Also, as β is close to 1, the γ parameter does not affect the curve, then reduced to a 2-dimensional curve function of α. This is consistent with formula (5) where the higher the value of α, the lower the impact of the two other parameters, and the higher the value of β the lower the impact of γ).

In average over all sequences the peak performance is reached for $\alpha = 0.42, \beta = 0.32, \gamma = 0.24$. The results for specific sequences are shown in table 1. We can see that the optimal weighting is very consistent over the different sequences, even when the quality

Sequence #	Chance	Cross-correlation	Multimodal	Weights of the peak performance
1	20%	26.0%	28.6%	$\alpha = 0.5, \beta = 0.4, \gamma = 0.2$
2	20%	45.0%	46.0%	$\alpha = 0.4, \beta = 0.4, \gamma = 0.3$
3	20%	56.2%	55.5%	$\alpha = 0.3, \beta = 0.2, \gamma = 0.2$
4	20%	68.2%	68.4%	$\alpha = 0.4, \beta = 0.2, \gamma = 0.3$
5	20%	65.3%	63.7%	$\alpha = 0.5, \beta = 0.4, \gamma = 0.2$
All	20%	52.1%	52.4%	$\alpha = 0.42, \beta = 0.32, \gamma = 0.24$

Table 1: Optimal parameters for each sequence, and comparison of performances.

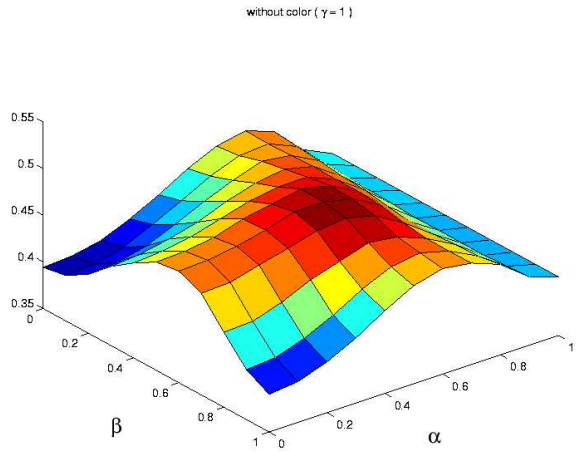


Figure 6: Graph of the disparity quality for different modality weights over all sequences, excluding the colour ($\gamma = 1$).

of the disparity changes drastically. The peak performance is reached for a strong use of the optic flow information ($\alpha \simeq 0.4$), showing the relevance of the optic flow modality for this task. Also, the algorithm performs slightly better than the cross correlation while *using only ten parameters instead of 300*.

4.3 Performance without Colour or Optic Flow

The performances with grey level images (γ is then set to 1) is shown in figure 6 and table 2, third column. Again the peak performance is reached for a significant use of optic flow. The peak performance drops by 2.7% compared to colour images and again by 2.5% if the optic flow is neglected ($\alpha = 0$). As expected, on figure 6 the performance decreases considerably when using only one of the modalities. This shows the relevance of this multimodal matching, and more specifically of the use of optic flow for this task.

In table 2, fourth column, is shown the performance of the program on colour images without using the optic flow information (*i.e.* with the parameter α set to 0). This respresents a drop in peak performance of 0.7%. Compared to the 2.5% with greyscale images, this leads us to assume that the use of different modalities improves the robustness as well as the general performance of the method. The marginal loss of performance when ignoring one of the most weighted modalities (2.7 percents for colour, and 0.7% for the optic flow, compared to the 5.2% of loss when neglecting those two), also confirms the

Sequence #	Chance	Multimodal without Colour	Multimodal without Optic Flow
1	20%	25.8%	26.7%
2	20%	43.3%	45.5%
3	20%	53.1%	54.6%
4	20%	64.5%	67.9%
5	20%	62.0%	63.6%
All	20%	49.7%	51.7%

Table 2: Performances of our function when excluding one parameter (colour or optic flow).

robustness of this multimodal similarity function.

5 Conclusion

In this paper we presented a multimodal similarity function and applied it to the stereo correspondance problem. We applied this method to several scenes of diverse difficulty and compared its performance with a standard normalized cross-correlation algorithm. The results clearly shown the importance of the optic flow in this method. It is also interesting to note that our data processing allows an important data reduction: this representation features *only ten parameters* (or 4 without the colour information) *instead of 300 for the cross correlation (100 without colour)*, which comes to a reduction of 96.6% (96% without the colour). In spite of this considerable condensation, we assume that no crucial information loss (relatively to the task) had happened, as the result matches the performances of the cross correlation, and even outperform it slightly on difficult scenes (emphasizing again the importance of the added optic flow information). The robustness of the method is outlined by the consistency of the optimal weights found over all sequences, while the peak performance itself varied largely..

Acknowledgements We would like to thank Florentin Wörgötter for fruitfull discussions.This project has been funded by ECOVISION.

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] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1971.

[4] A. Cozzi and F. Wörgötter. Comvis: A communication framework for computer vision. *International Journal of Computer Vision*, 41:183–194, 2001.

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

[6] O.D. Faugeras. *Three–Dimensional Computer Vision*. MIT Press, 1993.

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

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

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

[10] 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.

[11] R. Koch. Model-based 3-D scene analysis from stereoscopic image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 49(5):23–30, 1994.

[12] A. Koschan. Chromatic block matching for dense stereo correpondence. *Proceedings of ICIAP*, 1993.

[13] A. Koschan. How to utilize color information in dense stereo matching and in edge based stereo matching? *Proceedings of ICARCV*, pages 419–423, 1994.

[14] P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.

[15] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 15:111–118, 2002.

[16] N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. *BMVC2003*, 2003.

[17] N. Krüger, M. Felsberg, C. Gebken, and M. Pörksen. An explicit and compact coding of geometric and structural information applied to stereo processing. *Proceedings of the workshop 'Vision, Modeling and VISUALIZATION 2002'*, 2002.

[18] N. Krüger, M. Felsberg, and F. Wörgötter. Processing multi-modal primitives from image sequences. *submitted to BMVC 2003*, 2003.

[19] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *'AISB 2003 Convention: Cognition in Machines and Animals', Wales*, 2003.

[20] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *Network: Computation in Neural Systems*, 13:553–576, 2002.

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

[22] H.-H. Nagel. On the estimation of optic fbw: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[23] S.J. Sangwine and R.E.N. Horne. *The Colour Image Processing Handbook*. Chapman & Hall, 1998.

[24] C. Schmid and A. Zisserman. Automatic line matching across views. *Proc. IEEE Conference on Computer Vision and Pattern Recognition,*, pages 666–671, 1997.

[25] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, London, 1999.