Empirically Convergent Adaptive Estimation of Grayvalue Structure Tensors

Markus Middendorf and Hans-Hellmut Nagel

Institut für Algorithmen und Kognitive Systeme,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany
Phone: +49-721-608-4044/4323
Fax: +49-721-608-8481
markusm|nagel@ira.uka.de

Abstract. An iterative adaptation process for the estimation of a Grayvalue Structure Tensor (GST) is studied experimentally: alternative adaptation rules, different parameterizations, and two convergence criteria are compared. The basic adaptation process converges for both synthetic and real image sequences in most cases towards essentially the same results even if different parameters are used. Only two identifiable local grayvalue configurations have been encountered so far where adaptation variants do not converge according to the chosen criteria.

Keywords: Image features, Visual motion, Adaptive estimation, Grayvalue Structure Tensor, Anisotropic filtering, Optical flow.

1 Introduction

About ten years ago, Koenderink and van Doorn studied a family of generic *neighborhood* operators [6], taking into account aspects of many linear operators encountered for the extraction of *local* image properties. Based on what these authors take to be fundamental hypotheses, they consider a multivariate isotropic Gaussian kernel as the most elementary linear 'neighborhood' operator: it estimates the illuminance at each 'point' by a weighted average of grayvalues around the specified image position.

The *extent of the neighborhood* is quantified by the standard deviation of the Gaussian kernel. This standard deviation can be taken to represent the 'scale' of image properties considered to be relevant for a particular investigation of the illuminance distribution impinging onto the image plane. The *internal structure* of the grayvalue variation within such a neighborhood can be characterized by (functions of) more detailed local estimates obtained by the convolution of the image with partial derivatives of the Gaussian kernel.

Previous work, e.g. [11], shows, that it may be useful *not to fix a scale globally*, but to estimate the relevant scale in each dimension *locally* for each neighborhood, and, that the 'most appropriate scales' do not need to occur exactly along the axes of the initially chosen coordinate system. This procedure implies a 'hen-and-egg' problem: The neighborhood is specified by a chosen scale, whereas, at the same time, the scale has to be determined from the neighborhood properties.

This contribution studies a particular approach – the *adaptive estimation of a Grayvalue Structure Tensor (GST)* – in order to escape from this dilemma. The goal is to provide evidence that one can 'operationalize' the concept of an 'anisotropic neighborhood' in a *spatiotemporal* grayvalue distribution, provided certain boundary conditions are accepted.

2 On Related Publications

Since many years, 'anisotropic diffusion' constitutes an active research area – see, e.g., [17]. Originally, it has been studied for image enhancement by anisotropic smoothing and for edge extraction – see, e.g., [14, 7]. Similar adaptive approaches have been used, too, in order to increase the robustness of structure extraction in cases where neighboring line structures have to be resolved, e. g., for automatic analysis of fingerprint images [1]. An early application of such ideas to Optical Flow (OF) estimation has been discussed in [4] – see, too, [5, Chapter 10 'Adaptive Filtering']. A kind of closed-form adaptive GST estimation approach for the estimation of OF-fields has been reported in [11], a more detailed study of adaptation effects for the same purpose in [12].

Recently, the aspect to reduce the influence of noise by a variant of anisotropic diffusion has been studied specifically in the context of GST-based OF estimation [15] where one can find additional references regarding this topic. Earlier publications about OF estimation have been surveyed in [3].

Usually, only one or two iterations are used for adaptation (see, too, [9, 10]), although an adaptation algorithm 'naturally' raises a question regarding its convergence and its dependence on initial conditions or other parameters.

3 Adaptive Estimation of the GST

The GST is a weighted average (over a neighborhood of the current pixel) of the outer product of $\nabla g = (\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}, \frac{\partial g}{\partial t})^T$ with itself, where ∇g denotes the gradient of the greyvalue function $g(x, y, t)$ with respect to image plane coordinates (x, y) and time t. In both gradient computation and averageing, the extent of the Gaussian kernels involved is determined by their covariances Σ_{G} and Σ_{A} respectively. Inspired by [8], we use $\Sigma_{\mathrm{A}} \simeq 2 \cdot \Sigma_{\mathrm{G}}$.

In what follows, we basically use the same algorithms as proposed by [10], with some differences that will be pointed out in the sequel:

During the location-invariant 'initialisation step', the same fixed 'start covariance' matrices $\Sigma_{\mathrm{G}0} = \mathrm{diag}(\sigma_{xy}, \sigma_{xy}, \sigma_t)$ and $\Sigma_{\mathrm{A}0}$ are used at every pixel position to determine GST_0 ("0-th iteration").

The subsequent GST *estimation phase adapts* Σ_{G} and thus Σ_{A} iteratively to the prevailing local grayvalue variation. During the i-th iteration ($i \geq 1$),

$\Sigma_{\mathrm{G}i}$ is basically set to the *inverse* of GST_{i-1}. GST_{i-1} is decomposed into a rotational matrix U and a diagonal matrix $E = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$, comprising the eigenvalues $\lambda_i, i = 1, 2, 3$. The extent of the Gaussian kernel is then derived from E, its orientation from U. In contrast to [10], we shall compare two alternatives to compute the kernel's extent, namely the 'linear approach' proposed in [12]

$$\alpha_k = \tilde{\lambda}_k \cdot {\sigma_{\min}}^2 + (1 - \tilde{\lambda}_k) \cdot {\sigma_{\max}}^2 \qquad (1)$$

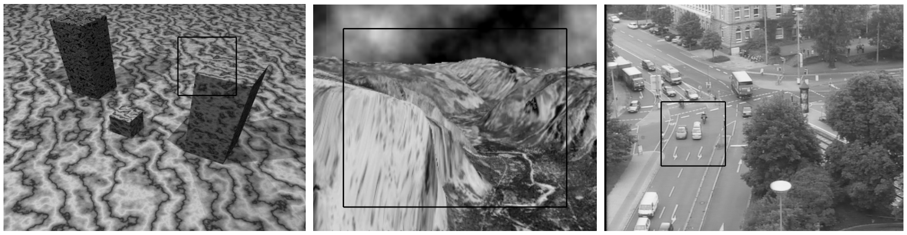
and the 'inversion approach' first discussed by [10]

$$\alpha_k = \frac{{\sigma_{\max}}^2 \cdot {\sigma_{\min}}^2}{(1 - \tilde{\lambda}_k) \cdot {\sigma_{\min}}^2 + \tilde{\lambda}_k \cdot {\sigma_{\max}}^2} \quad , \qquad (2)$$

where $\tilde{\lambda}_k$ are the 'scaled eigenvalues' $\lambda_k / (\lambda_1 + \lambda_2 + \lambda_3)$, and α_k is the extent in the direction of the k-th eigenvector.

4 Experimental Setup

Although convergence is a well studied problem in mathematics, only experience can show whether the approach outlined above will converge under the additional constraints set by the algorithm, e.g. finite mask size, and – if so – which iteration alternatives and parameter settings lead to the desired result. We evaluated three image sequences with different characteristics (see Figure 1): A rendered, noise-less MARBLE BLOCK scene with known shift rate, used to check the implementation; the well known YOSEMITE sequence ([19]); and a real-world INTERSECTION scene ([18]), used to evaluate the algorithm's behaviour on noisy data.



MARBLE BLOCK scene YOSEMITE scene INTERSECTION scene

Fig. 1. Representative frames from the image sequences studied here (the rectangles mark the clipped image area used from each sequence).

The following questions are investigated in this context: Will the iterative approach (or at least some of the alternatives studied) 'converge' at all? Does the resulting GST-estimate depend on the initial conditions? And finally, which grayvalue configurations in an image sequence result in 'fast' convergence and which ones present greater obstacles?

An iterative GST estimation will be taken to converge if 'some scalar quantification' of the difference between GST_i and GST_{i-1} drops below a threshold.

A straightforward approach consists in using the Frobenius norm $\|\mathrm{GST}_i - \mathrm{GST}_{i-1}\|$, in our case with a threshold of 0.03 which corresponds to an average error of 0.01 in each matrix component.

In order to sharpen our intuition about what happens in case the iteration converges slowly or even fails to converge, we have chosen in addition an alternative, indirect convergence criterion, namely convergence of OF estimation associated with the GST. An OF-vector is treated as a three-dimensional spatiotemporal vector $\boldsymbol{u} = (u_1, u_2, 1)^T$. Based on a 'Total Least Squares (TLS)' approach, the OF-vector $\boldsymbol{u}(\boldsymbol{x})$ at position $\boldsymbol{x}$ will be determined from the eigenvector $\boldsymbol{e}_{\min}$ corresponding to the smallest eigenvalue $\lambda_{\min}$ of the GST (see, e. g., [16]). We consider the iterative GST estimation as convergent if the difference $\|\boldsymbol{u}_i(\boldsymbol{x}) - \boldsymbol{u}_{i-1}(\boldsymbol{x})\|_2$ between consecutive OF-estimates at image location $\boldsymbol{x}$ becomes smaller than $^1\!/_{1000}$ pixel per frame (ppf). Having in mind that ground truth of, e. g., the YOSEMITE sequence exhibits already a *quantisation uncertainty* of 0.0316 ppf, a convergence threshold of less than 0.0010 ppf appears acceptable.

Evidently, quantification of convergence based on the OF-difference and based on $\|\mathrm{GST}_i - \mathrm{GST}_{i-1}\|$ need not lead to convergence after the same number of iterations or may even yield different results regarding convergence in certain cases. Our experimental study showed, however, that both criteria are more or less equivalent.

5 Experimental Results

In a first series of experiments, we computed 100 adaption steps at each pixel position for each image sequence, thresholding the difference between consecutive OF-estimates as convergence criterion. The following alternatives have been compared:

- two different adaptation algorithms: linear vs. inversion (see Section 3);
- two different start covariances: $\sigma_{xy} = \sigma_t = 1.0$ vs. $\sigma_{xy} = 2.0, \sigma_t = 1.0$, and
- two different adaptation areas: $\left({\sigma_{\min}}^2 = 1.0 \text{ pixel}^2 \text{ and } {\sigma_{\max}}^2 = 8.0 \text{ pixel}^2\right)$ vs. $\left({\sigma_{\min}}^2 = 1.0 \text{ pixel}^2 \text{ and } {\sigma_{\max}}^2 = 16.0 \text{ pixel}^2\right)$;

except for the YOSEMITE sequence where we omitted the run with the large adaptation area due to the limited number (15) of frames, which would otherwise reduce the temporal extent of the convolution masks in an improper way.

Figure 2 plots the percentage of pixels for each of the three clippings as a function of the iteration count after convergence, when using the 'inversion' approach. The convergence threshold serves as curve parameter. The results for the 'linear' approach differ only by a few percent; the difference would hardly be visible in a plot like the one above. The results for the comparison between different *start* covariances are similar; neither of the two parameter sets exhibits obvious advantages.

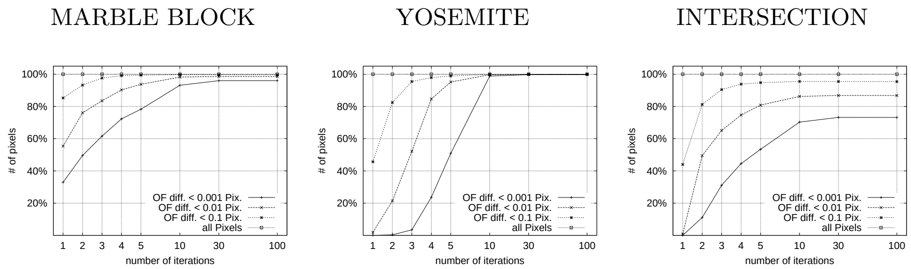


Fig. 2. Convergence for the ‘inversion’ approach, depending on convergence threshold.

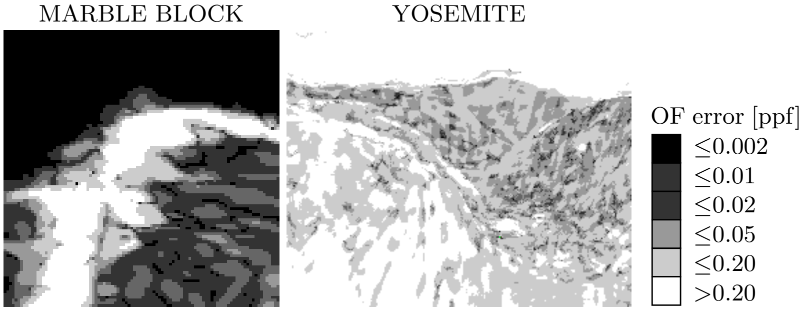


Fig. 3. Comparison of estimated Optical Flow with known ground truth. Experimental setup: inversion approach, iteration # 10, $\Sigma_{\mathrm{G}} = \mathrm{diag}(2.0, 2.0, 1.0)$, ${\sigma_{\max}}^2 = 8.0$ pixel^2.

The only significantly different results were observed when varying the range of admitted adaptation: For a larger adaptation area, the convergence rate was smaller. This observation is not surprising: a larger range of adaption leads to a higher number of configurations where convergence is prevented by special conditions (see later in this section).

As pointed out above, the determination of the eigenvector associated with the smallest eigenvalue of a convergent GST estimate provides a basis to estimate an OF-vector. Figure 3 compares estimates obtained in this manner for the two image sequences MARBLE BLOCK and YOSEMITE where the ‘ground truth’ is known: the color-coded difference between the ‘true’ value and the estimated one is superimposed to the clipping from the original image. As one can see immediately, discrepancies in the MARBLE BLOCK sequence are restricted to boundary curves where one expects discontinuities in the OF-field. In the YOSEMITE sequence, the discrepancies are much larger, but due to the quantization in the ground truth data, an average absolute error of about 0.02 ppf is to be expected even for ’perfectly’ estimated OF vectors. The assumption of brightness constancy does not hold for the clouds, thus the large estimation error there is not surprising. The errors in the lower left part of the image may be explained by the restrictions on the mask size, which affect especially these areas with a high shift rate of up to 4.9 ppf.

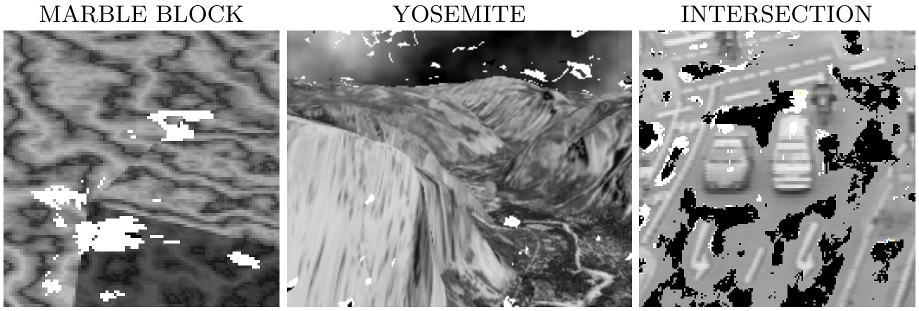


Fig. 4. Non-convergent cases after 10 iterations. The pixels are coloured according to their categorization: black – lack of image structure, results significantly influenced by noise; white – other categories. Experimental setup: inversion approach, $\Sigma_{\mathrm{G}} = \mathrm{diag}(2.0, 2.0, 1.0)$, ${\sigma_{\max}}^2 = 8.0$ pixel^2.

Although the results on the YOSEMITE sequence do not look very good at first sight, they are better then almost every other OF estimation method so far. The average angular error observed in our experiments ranges from 3.16° (at 49.4% density) to 8.40° (at 71.6% density), depending on the parameterization, which is better than all results reported by [3]. Even newer approaches, e.g. [13], give worse results. The only approach – as far as we know – with comparable results was suggested by Alvarez et al. ([2]), which produces a slighly higher average error, but with a density of 100%.

In general, convergence is observed after 10 or at most 30 iterations. Pixel positions without convergence even after 70 additional iterations were scrutinized in search for identifiable grayvalue structures which could possibly explain the lack of convergence. Figure 4 illustrates the distribution of pixels without convergence after 10 adaptive iterations. Whereas most non-convergent estimates occur along the edges of the moving block in the MARBLE BLOCK sequence, non-convergent cases in the INTERSECTION sequence occur in the image background. These different distributions may be roughly assigned to one of two characteristic grayvalue configurations to be discussed in the sequel.

In the INTERSECTION sequence, most non-convergent pixels can be found in the image background (see right column of Fig. 4). In most cases (illustrated here by pixels painted black), trace(GST) is smaller than 5.0, indicating a lack of image structure: shape and extent of masks after adaptation are significantly influenced by noise. Small variations of the masks' extent may lead to significantly different derivatives, thus resulting in a different OF-estimate. Areas with low image structure occur only in the INTERSECTION sequence, and – consistently – non-convergent pixels with small trace(GST) are only observed in that sequence.

The MARBLE BLOCK sequence comprises only textured surfaces. Thus lack of image structure can not serve as an explanation for convergence failures.

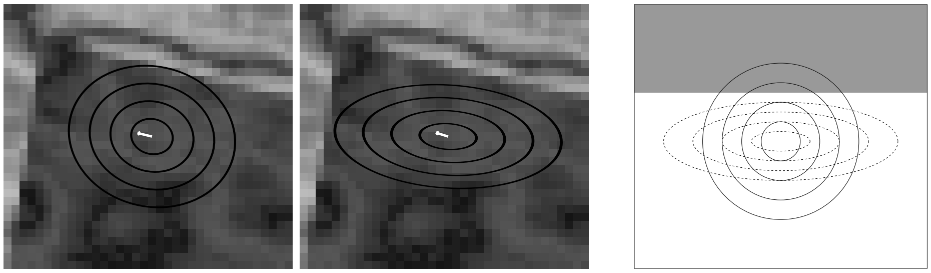

iteration 18 iteration 19

Fig. 5. Left and middle panel: Extent of the convolution masks, projected into the image plane. The black lines show the intersection of the ellipsoids containing 1, 2, 3, or 4 standard deviations, respectively, of a Gaussian with covariance matrix Σ_A. The example presented here shows one of those pixel locations where 'oscillation' can be observed most clearly. Right panel: Schematic explanation for the effect illustrated here. The initial mask (solid lines) with equal extents in both directions touches an edge. During adaptation, the mask is thus extended along the edge and compressed perpendicular to it. The resulting new mask (dashed lines) does no longer cover the edge, thus resulting – during the second iteration – in a mask with the same extent as the original, unadapted mask.

Figure 5 shows a typical example for non-convergence in the MARBLE BLOCK sequence. After a few iterations, the mask extent oscillates between two states. The difference between iteration 18 and 19 – and between iteration 19 and 20 – is about $^{1}/_{2}$ ppf, whereas the OF-estimates differ only minimally ($\approx {}^{1}/_{1000}$ ppf) between iteration 18 and 20.

6 Summary and Conclusions

It appears as a remarkable result that iterative adaptive GST-estimation converges at most image positions for all combinations of algorithmic alternatives, parameter selections, and test image sequences investigated. The degree of convergence differs somewhat in detail, depending on the image data: the convergence results for YOSEMITE (≈ 99% after 10 iterations) and MARBLE BLOCK (≈ 95% after 10 it.) are better than for INTERSECTION (≈ 80% after 10 it.). Given significant image areas in the INTERSECTION sequence with only minor grayvalue variations, but noticeable noise, this outcome does not really surprise. Generally, convergence seems to depend more on the properties of the image sequence than on the choice of parameter.

The right panel of Figure 5 explains the results in a simplified example where the masks in subsequent iterations alternately include and exclude an edge. The case in the left and middle panel is even more complicated: The elongated masks in iterations 17 and 19 include the edge on the lefthand side, but exclude the edge above, whereas in iteration 18, the edge above the pixel is covered and the edge on the left is excluded.

Based on the experience accumulated so far, our approach either converges towards a description of dominant local characteristics which can be represented by an anisotropic spatiotemporal 'Gaussian bell' or it points towards an *identifiable* grayvalue configuration which is incompatible with this representation.

Acknowledgements

Partial support by the Deutsche Forschungsgemeinschaft and the European Union (IST-2000-29404) for the investigations reported here is gratefully acknowledged.

References

1. A. Almansa and T. Lindeberg: Fingerprint Enhancement by Shape Adaptation of Scale-Space Operators with Automatic Scale Selection. IEEE Tr. on Image Processing **9**:12 (2000) 2027–2042.
2. L. Alvarez, J. Weickert, and J. Sanchez: Reliable estimation of dense optical flow fields with large displacement. International Journal on Computer Vision **39**:1 (2000) 41–56.
3. J. Barron, D. Fleet, and S. Beauchemin: Performance of Optical Flow Techniques. International Journal on Computer Vision **12**:1 (1994) 156–182.
4. J. Bigün, G.H. Granlund, and J. Wiklund: Multidimensional Orientation Estimation with Applications to Texture Analysis and Optical Flow. IEEE Tr. on Pattern Analysis and Machine Intelligence PAMI-**13**:8 (1991) 775–790.
5. G.H. Granlund and H. Knutsson: Signal Processing for Computer Vision. Kluwer Academic Publishers, Dordrecht Boston London 1995.
6. J.J. Koenderink and A. van Doorn: "Generic Neighborhood Operators". IEEE Tr. on Pattern Analysis and Machine Intelligence PAMI-**14**:6 (1992) 597-605.
7. T. Leung and J. Malik: Contour continuity in Region Based Image Segmentation. Proc. ECCV 1998, Vol. I, H. Burkhardt and B. Neumann (Eds.), LNCS 1406, pp. 544–559.
8. T. Lindeberg and J. Gårding: "Shape-Adapted Smoothing in Estimation of 3-D Depth Cues from Affine Distortions of Local 2-D Brightness Structure". Proc. ECCV 1994, J. O. Eklundh (Ed.), LNCS 800, pp. 389–400.
9. M. Middendorf and H.-H. Nagel: Vehicle Tracking Using Adaptive Optical Flow Estimation. Proc. First Int. Workshop on Performance Evaluation and Surveillance (PETS 2000), 31 March 2000, Grenoble, France. J. Ferryman (Ed.), The University of Reading, Reading, UK, 2000.
10. M. Middendorf and H.-H. Nagel: Estimation and Interpretation of Discontinuities in Optical Flow Fields. Proc. ICCV 2001, 9-12 July 2001, Vancouver/Canada; Vol. I, pp. 178–183.
11. H.-H. Nagel, A. Gehrke, M. Haag, and M. Otte: Space- and Time-Variant Estimation Approaches and the Segmentation of the Resulting Optical Flow Fields. Proc. ACCV 1995, 5-8 Dec. 1995, Singapore; In S.Z. Li, D.P. Mital, E.K. Teoh, and H. Wang (Eds.), Recent Developments in Computer Vision, LNCS 1035, pp. 81-90.
12. H.-H. Nagel and A. Gehrke: Spatiotemporally Adaptive Estimation and Segmentation of OF-Fields. Proc. ECCV 1998, 2-6 June 1998, Freiburg/Germany; H. Burkhardt and B. Neumann (Eds.), LNCS 1407 (Vol. II), pp. 86–102.
13. K. P. Pedersen and M. Nielsen: Computing Optic Flow by Scale-Space Integration of Normal Flow. In M. Kerckhove (Ed.) "Scale-Space and Morphology in Computer Vision", Proc. 3rd International Conference on Scale-Space, LNCS 2106, 2001.
14. P. Perona and J. Malik: Scale Space and Edge Detection Using Anisotropic Diffusion. IEEE Tr. on Pattern Analysis and Machine Intelligence PAMI-**12**:7 (1990) 629-639.
15. H. Spies and H. Scharr: Accurate Optical Flow in Noisy Image Sequences. In Proc. 8th International Conference on Computer Vision ICCV 2001, 9–12 July 2001, Vancouver, BC, Vol. I, pp. 587–592.
16. J. Weber and J. Malik: "Robust Computation of Optical Flow in a Multi-Scale Differential Framework.". Proc. ICCV-93, pp. 12–20; see, too, Intern. Journ. Computer Vision **14**:1 (1995) 67–81.
17. J. Weickert: Anisotropic Diffusion in Image Processing. B.G. Teubner, Stuttgart, Germany 1998.
18. Universität Karlsruhe: `http://i21www.ira.uka.de/image_sequences/` – the INTERSECTION scene is called "Karl-Wilhelm-Straße (normal conditions)" there.
19. Originally produced by Lynn Quam at SRI, available at the University of Western Ontario: `ftp://ftp.csd.uwo.ca/pub/vision/TESTDATA/YOSEMITE_DATA/`.