Proc. 5th International Conference on Computer Vision, (Cambridge, MA), June 1995, pages 134–141.

Direct estimation of affine image deformations using visual front-end operations with automatic scale selection

Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)*
KTH (Royal Institute of Technology), Stockholm, Sweden

Abstract: This article deals with the problem of estimating deformations of brightness patterns using visual front-end operations. Estimating such deformations constitutes an important subtask in several computer vision problems relating to image correspondence and shape estimation. The following subjects are treated:

The problem of decomposing affine flow fields into simpler components is analysed in detail. A canonical parametrization is presented based on singular value decomposition, which naturally separates the rotationally invariant components of the flow field from the rotationally variant ones.

A novel mechanism is presented for automatic selection of scale levels when estimating local affine deformations. This mechanism is expressed within a multi-scale framework where disparity estimates are computed in a hierarchical coarse-to-fine manner and corrected using iterative techniques. Then, deformation estimates are selected from the scales that minimize a certain normalized residual over scales. Finally, the descriptors so obtained serve as initial data for computing refined estimates of the local deformations.

1 Introduction

In several computational vision models, the deformations of brightness patterns constitute an important modelling step. When a camera fixates a surface pattern in the world, the pattern is deformed when mapped onto the camera by the perspective transformation. The structure of this deformation is determined both by the shape of the object and the orientation of the object relative to the observer. In terms of this framework a large number of visual modules can be expressed, such as motion estimation, structure from motion, stereo matching, vergence control, shape estimation from binocular data, shape from texture, etc. In general, these deformations can be modelled by projective transformations. Approximating the models by local first-order approximations (derivatives) gives rise to affine transformations.

This article deals with the problem of measuring such local transformations between two-dimensional images. Whereas this problem can be and has been studied in the contexts of specific shape-from-X competences and using the geometric information available in any specific case, it is important to study the general problem of estimating image deformations based on two-dimensional image information only. One reason is the generality of the approach and the potential in expressing different shape-from-X competences using a similar theoretical framework and similar image operations. (Thereby decoupling specific geometric information or assumptions from image measurements.) Another motivation is that disregarding oculomotoric cues, this is the only information available to an uncommitted vision system without specific knowledge about the world.

To simplify the presentation, we shall throughout consider the specific case with only two images, corresponding to binocular stereo. When analysing motion data, it is, of course, generally agreed upon that better performance can be obtained by studying coherent data over time than just two single time moments. In that case, we assume that the raw spatio-temporal data have already been pre-processed in a spatio-temporal scale-space representation comprising averaging over both space and time. The image pairs to this analysis will then be image slices from adjacent time moments at some temporal scale. This is in analogy with the situation in the regular (spatial) scale-space representation, where nearest-neighbourhood operations are known to be highly noise sensitive at the finest levels of scale, but nevertheless deliver highly useful and robust results when applied at sufficiently coarse scales.

Because of the generality of this problem domain, these problems have been extensively studied in the literature, and it is impossible to make a fair review here. Besides the explicit citations here, the reader is referred to the recent overview by (Barron *et al.* 1994) and a longer version of this manuscript (Lindeberg 1994b).

*This work was partially performed under the Esprit-BRA project InSight and the Esprit-NSF collaboration Diffusion. The support from the Swedish Research Council for Engineering Sciences, TFR, is gratefully acknowledged.
Address: NADA, KTH, S-100 44 Stockholm, Sweden
Email: tony@bion.kth.se (http://www.bion.kth.se/~tony).

The presentation is organized as follows: We first analyse in detail the problem of decomposing and parametrizing affine transformations. Then, we turn to the problem of estimating these deformations.

2 Affine image transformations

An affine image transformation of a point $x \in \mathbb{R}^N$ to a new position $x' \in \mathbb{R}^N$ can be represented by

$$x' = \mathcal{A}x + b. \tag{1}$$

This transformation arises, for example, as the result of truncating all terms of higher order than one in the Taylor expansion of a general spatial transformation $x' = \mathcal{T}(x)$. Here, shall be throughout concerned with the two-dimensional case. With $x = (x_1, x_2)^T$ and $x' = (x_1', x_2')^T$, the explicit coordinate representation is

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{2}$$

A general assumption we make is that the deformations are small, *i.e.* that the matrix $\mathcal{A}$ is close to the identity matrix. In particular, we can hence exclude degenerate transformations as well as reflections.

2.1 Classification based on the eigenvalues of $\mathcal{A}$

If (1) is used for iterative movement of points,

$$x^{(k+1)} = \mathcal{A}\, x^{(k)} + b, \tag{3}$$

then a discrete flow field is generated. This flow field can be interpreted as a unit time step discretization of the corresponding differential equation

$$\dot{x}(t) = (\mathcal{A} - \mathcal{I})\, x(t) + b, \tag{4}$$

where $\mathcal{I}$ denotes the identity matrix. Depending on the eigenvalues λ_1 and λ_2 of $\mathcal{A}$, qualitatively different types of flow fields can be distinguished (see table 1 for an illustration). In this respect, the eigenvalues of $\mathcal{A}$ provide a taxonomy for classifying affine flow fields.

Type	Eigenvalues	Representative $\mathcal{A}$
Expansion Contraction Saddle	$\lambda_1 > \lambda_2 > 1$ $\lambda_1 < \lambda_2 < 1$ $\lambda_1 > 1, \lambda_2 < 1$	$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$
Jordan	$\lambda_1 = \lambda_2$ real	$\begin{pmatrix} \lambda & -\lambda \tan\varphi \\ 0 & \lambda \end{pmatrix}$
Rotation	non-real: $\rho e^{\pm i\phi}$	$\rho \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$

Table 1: Examples of characteristic affine flow fields arising from a classification based on the eigenvalues of $\mathcal{A}$.

3 Parametrizing affine transformations

A classification of flow fields in terms of the eigenvalues of $\mathcal{A}$, however, reflects only the linear structure of the transformation. In geometric problems, where a metric structure is present as well, such as orthogonality and distances, singular value decomposition is a more powerful tool for expressing linear transformations.

This section shows how a canonical representation of two-dimensional retinal flow fields can be introduced based on this idea. The resulting representation is closely related to the div–curl–def descriptors introduced by (Koenderink and van Doorn 1975). An advantage of the proposed parametrization, however, is that the singular value decomposition completely reveals the structure of the affine transformations. In particular, it makes the distinction more explicit between the two different cases when the relative torsion states of two cameras are either known or unknown. In certain literature, these notions have been confused.

3.1 Rotationally invariant descriptors of $\mathcal{A}$

Consider the effect of performing arbitrary rotations of the domains where x and x' in (1) are defined: Let

$$u' = \mathcal{R}_\alpha\, x' \quad \text{and} \quad u = \mathcal{R}_\beta x, \tag{5}$$

where $\mathcal{R}_\alpha$ and $\mathcal{R}_\beta$ represent rotations by angles α and β in the counter-clockwise direction respectively

$$\mathcal{R}_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \quad \mathcal{R}_\beta = \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix}.$$

For $u' = \mathcal{A}'u + b'$ to hold, $\mathcal{A}$ and b must transform as $\mathcal{A}' = \mathcal{R}_\alpha\, \mathcal{A}\, \mathcal{R}_{-\beta}$ and $b' = \mathcal{R}_\alpha\, b$. From $a_{i,j}$ introduce

$$\begin{aligned} T &= (a_{11} + a_{22})/2, & A &= (a_{21} - a_{12})/2, \\ C &= (a_{11} - a_{22})/2, & S &= (a_{12} + a_{21})/2. \end{aligned} \tag{6}$$

Then, these descriptors transform according to

$$\begin{pmatrix} T' \\ A' \end{pmatrix} = \begin{pmatrix} \cos(\alpha-\beta) & -\sin(\alpha-\beta) \\ \sin(\alpha-\beta) & \cos(\alpha-\beta) \end{pmatrix} \begin{pmatrix} T \\ A \end{pmatrix}$$
$$\begin{pmatrix} C' \\ S' \end{pmatrix} = \begin{pmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{pmatrix} \begin{pmatrix} C \\ S \end{pmatrix},$$

which corresponds to rotating $(T, A)^T$ and $(C, S)^T$ by angles $\alpha - \beta$ and $\alpha + \beta$. In particular, the descriptors

$$P^2 = T^2 + A^2, \quad \text{and} \quad Q^2 = C^2 + S^2 \tag{7}$$

are unaffected by rotations. In the special case when the rotations are performed symmetrically, *i.e.* $\alpha = \beta$, also $T' = T$ and $A' = A$ are rotationally invariant.

This special case is relevant, for example, when considering a flow field over time in a given coordinate system (*e.g.*, motion seen from a single camera) or between different coordinate systems for which the relative torsion states are known (*e.g.* calibrated stereo).

3.2 Parameters from singular value decomposition

The singular value decomposition of $\mathcal{A}$ is defined by

$$\mathcal{A} = \mathcal{U}\,\Sigma\,\mathcal{V}^T, \tag{8}$$

where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal matrices and Σ is a diagonal matrix. In the general case, $\mathcal{U}$ and $\mathcal{V}$ are not guaranteed to represent rotations, since orthogonal matrices also comprise reflections. Since the deformations are assumed to be close to the identity transformation, however, we can *require* $\mathcal{U}$ and $\mathcal{V}$ to represent rotations, $\mathcal{U} = \mathcal{R}_\alpha$ and $\mathcal{V} = \mathcal{R}_\beta$. Then, in the general case, Σ is not guaranteed to be a diagonal matrix with positive diagonal elements. For small deformations, however, that will be the case, and

$$\mathcal{A} = \mathcal{R}_\alpha\,\Sigma\,\mathcal{R}_{-\beta} \tag{9}$$

with $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2)$ and $\sigma_1, \sigma_2 > 0$. When expressed in terms of the *TACS* coordinates and the derived *PQ* entities, the closed form expression for the singular value decomposition is particularly simple. It is straightforward to verify that

$$\sigma_1 = P + Q, \quad \tan(\alpha - \beta) = A/T, \tag{10}$$

$$\sigma_2 = P - Q, \quad \tan(\alpha + \beta) = S/C, \tag{11}$$

and that the inverse relationships are

$$T = P\cos\theta, \quad A = P\sin\theta, \tag{12}$$

$$C = Q\cos\psi, \quad S = Q\sin\psi, \tag{13}$$

where the directional information is represented by

$$\theta = \alpha - \beta \quad \text{and} \quad \psi = \alpha + \beta. \tag{14}$$

In summary, this decomposition corresponds to

$$\mathcal{A} = \mathcal{R}_{\psi/2}\,\mathcal{R}_{\theta/2}\,\mathrm{diag}(\sigma_1, \sigma_2)\,\mathcal{R}_{\theta/2}\,\mathcal{R}_{-\psi/2}. \tag{15}$$

Alternatively, to obtain a maximally symmetric expression, we can rewrite the diagonal matrix as

$$\mathrm{diag}(\sigma_1, \sigma_2) = \sqrt{\sigma_1\sigma_2}\;\mathrm{diag}(\sqrt{\frac{\sigma_1}{\sigma_2}}, \sqrt{\frac{\sigma_2}{\sigma_1}}). \tag{16}$$

It is illuminating to compute these descriptors for the flow fields in table 1. In summary, the geometric interpretations of these entities are as follows:

- $\sigma_1\sigma_2 = P^2 - Q^2$ gives the *amount of expansion*.
- Q (or σ_1/σ_2) measures the *anisotropy* of the transformation. $Q = 0$ (or $\sigma_1/\sigma_2 = 1$) for transformations in the similarity group (translations, rotations, and uniform expansions/contractions).
- $\theta = \alpha - \beta$ reflects the *average amount of rotation*. $\theta = 0$ for expansions, contractions, saddles and translations. For rotations, θ is equal to the rotation angle, while for Jordan (skew) transformations, it is a trigonometric average of the maximally and minimally rotated directions.
- $\psi/2 = (\alpha + \beta)/2$ gives the direction of a *preferred symmetry axis* of the transformation. This symmetry axis is undetermined when $Q = 0$.

3.3 Summary and discussion

The singular value decomposition gives rise to a canonical decomposition and parametrization of small deformation affine flow fields, for which the rotationally invariant information in the singular values is completely decoupled from the rotationally dependent θ and ψ information. This property is important, for example, when computing affine transformations between images obtained from two metric cameras with unknown relative torsion states. If a calibration of the relative torsion states can be performed, then the information that can be extracted is perfectly captured by the singular values and θ. If on the other hand the cyclotorsion is unknown, σ_1 and σ_2 are the only invariant components.

Related representations. (Koenderink and van Doorn 1975) proposed a decomposition of motion flow fields in terms of three components called div, curl and def. Basically, these entities correspond to T, A and Q above, and to decomposing $\mathcal{A}$ into

$$\mathcal{A} = T\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + A\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} + Q\,\mathcal{M},$$

where $\mathcal{M}$ is a matrix containing directionally dependent information. As pointed out by (Koenderink and van Doorn 1975), the div, curl and def entities are unaffected by rotations of a common coordinate system ($\alpha = \beta$). Geometrically, this corresponds to the relative orientation states of the cameras being known (calibrated stereo) or the motion field registered from a camera in a fixed torsion state. In that case, the choice of primitives is, of course, arbitrary and the div–def–curl decomposition is functionally equivalent to the $P^2Q^2\theta\psi$ and $\sigma_1\sigma_2\theta\psi$ parametrizations. The advantage of the latter systems in these cases is that the transition to an unknown torsion state is a simple projection.

The decompositions induced by the *TACS* and the $PQ\theta\psi$ parameters also have the conceptual advantages that the *TACS* decomposition is purely linear and the $PQ\theta\psi$ decomposition is a pure matrix product. In this respect, the algebraic structure is cleaner.

A related representation for symmetric positive semidefinite matrices has been considered in (Lindeberg and Gårding 1993). In that representation, a *PCS* system is defined by $P = a_{11} + a_{22}$, $C = a_{11} - a_{22}$ and $S = 2a_{12} = 2a_{21}$. The main difference compared to the *TACS* system is that the symmetry requirements are relaxed and the effects of arbitrary rotations analysed.

Comparison with eigenvalue decomposition. Let us conclude this analysis by noting the difference between a singular value decomposition and a decomposition in terms of eigenvalues and eigenvectors. As remarked in the introduction, the eigenvalues and the eigenvectors depend only on the linear structure of the transformation and are as such independent of any metric. The singular value decomposition, on the other hand, is based on the existence of inner products and the notion of metric entities, such as distances and angles. If we are to capture the latter information, the singular value decomposition is the natural choice of these two.

4 Measuring affine transformations

Let us now turn to the problem of measuring image deformations. A common approach for stereo matching and computing three-dimensional shape cues has been to compute image features, such as points and lines, in an initial processing step, and then using these descriptors as primitives. Whereas a substantial simplification of the subsequent processing stages may be the result if reliable image features can extracted, the selection step crucially determines what results can be obtained and is often non-trivial. Therefore, it is of interest to consider methods that operate on the image intensities directly, using only filter-based operations and architecturally simple combinations of their outputs.

A fundamental problem in this context concerns what image operations to use. Is any operation feasible? A systematic approach that has been developed to restrict the class of possibilities is to assume that the first stages of visual processing should be as *uncommitted* as possible and have no particular bias. The essence of the results from scale-space theory (Witkin 1983; Koenderink and van Doorn 1990; Florack *et al.* 1992; Lindeberg 1994a) is that within the class of linear operations, convolution with Gaussian kernels and their derivatives is singled out as a canonical choice.

In this section, we shall consider a hierarchical differential flow field estimation approach closely related to (Bergen *et al.* 1992); see also (Werkhoven and Koenderink 1990; Jones and Malik 1992; Proesmans *et al.* 1994; Manmatha 1994; Sato and Cipolla 1994). We start by outlining a multi-scale disparity estimation framework that in addition to iterative corrections comprises bidirectional matching and explicit usage of confidence measures. Then, a scale selection mechanism is introduced based on the minimization of a certain normalized residual over scales. An attractive property of this approach is that the influence of disparity estimates at the finest scales is suppressed for noisy data that cannot be matched at fine scales.

4.1 Deformation measurements in scale-space

The scale-space representation L of a signal f is obtained by convolving f with Gaussian kernels $g(x;\, t) = 1/(2\pi t)\exp(-x^T x/2t)$ at different scales t. From this representation, Gaussian derivatives are defined by $L_{x^\alpha}(\cdot;\, t) = \partial_{x^\alpha} L(\cdot;\, t)$ where $\partial_{x^\alpha} = \partial_{x_1^{\alpha_1}} \partial_{x_2^{\alpha_2}}$.

Transformations in the similarity group. This representation is closed under transformations in the similarity group, *i.e.*, if two signals are related by $f_L(\xi) = f_R(\sigma \mathcal{R}_\varphi \xi + b)$, where $\mathcal{R}_\varphi$ is a rotation matrix, σ represents a positive scaling factor, and b a translation, the scale-space representations of f_L and f_R are related by $L(\xi;\, t) = R(\sigma \mathcal{R}_\varphi \xi + b;\, \sigma^2 t)$. Hence, for these transformations, the scale-space representations of f_L and f_R can always be perfectly matched.

Affine transformations and affine scale-space. To enable exact measurements of affine transformations with distinctly different singular values (*i.e.*, $Q \neq 0$), it is natural to generalize to non-symmetric Gaussians

$$g(x;\, \Sigma_t) = \frac{1}{2\pi\sqrt{\det \Sigma_t}} e^{-x^T \Sigma_t^{-1} x/2}, \tag{17}$$

whose shapes are controlled by covariance matrices Σ. For any function f the *affine Gaussian scale-space representation* (Lindeberg 1994a) L of f is defined as

$$L(\cdot;\, \Sigma_t) = g(\cdot;\, \Sigma_t) * f(\cdot). \tag{18}$$

Given two intensity patterns f_L and $f_R \colon \mathbb{R}^2 \to \mathbb{R}$ related by $f_L(\xi) = f_R(A\xi + b)$, the corresponding affine scale-space representations are related by

$$L(\xi;\, \Sigma_L) = R(A\xi + b;\, \Sigma_R) \quad \text{where} \quad \Sigma_R = A\, \Sigma_L A^T.$$

Compared to the non-linear affine invariant evolution schemes proposed by (Sapiro and Tannenbaum 1993; Alvarez *et al.* 1993) the advantage of this linear scale-space concept is that the scale-space properties transfer to all derivatives. The disadvantage is that it leads to a three-parameter variation.

4.2 Establishing correspondence

A fundamental problem when estimating image deformations concerns how to establish correspondence between different images of the same scene. Whereas the commonly used constant brightness assumption suffers from inherent limitations, we shall nevertheless use it for establishing an initial correspondence. (Then, it can be applied to other differential descriptors, such as the Laplacian.) Hence, assume that

$$f_R(\xi) = f_L(\xi + \Delta\xi) = f_L(\xi) + (\nabla f_L)(\xi)\,\Delta\xi + \mathcal{O}(|\Delta\xi|^2)$$

and consider only the first-order terms. This gives rise to (the discrete form of) the well-known motion constraint equation (Horn and Schunck 1981)

$$(\nabla f_L)(\xi)^T(\Delta\xi) + (f_L(\xi) - f_R(\xi)) = \mathcal{O}(|\Delta\xi|^2).$$

Since this analysis is compatible with brightness measurements in scale-space, at any scale t we also have

$$(\nabla L)(\xi;\, t)^T(\Delta\xi) + (L(\xi;\, t) - R(\xi;\, t)) = \mathcal{O}(|\Delta\xi|).$$

Least-squares estimation. Assume that the motion field can be approximated by a constant flow field v over the support region of a window function w. Following (Bergen *et al.* 1992; Barron *et al.* 1994) and several others, integrate the square of this relation using w as window function. After expansion (and dropping the arguments) this gives the least squares problem

$$\min_{v \in \mathbb{R}^2} v^T A v + 2b^T v + c, \tag{19}$$

where A, b, and c are defined by

$$A = \int_{\xi \in \mathbb{R}^2} (\nabla L)(\nabla L)^T\, w\, d\xi \tag{20}$$

$$b = \int_{\xi \in \mathbb{R}^2} (R - L)\,(\nabla L)\, w\, d\xi \tag{21}$$

$$c = \int_{\xi \in \mathbb{R}^2} (R - L)^2\, w\, d\xi. \tag{22}$$

Ambiguity. Of course, when treated pointwise, the motion constraint equation only determines the normal flow parallel to ∇L. If, however, the support region of w contains a sufficiently rich distribution of (coherently moving) gradient directions, the solution to (19) may give an estimate close to the true flow field. A natural measure of how scattered the gradient directions are is given by the normalized anisotropy (derived from A)

$$\tilde{Q} = Q/P. \tag{23}$$

When all gradient directions are parallel, we have $\tilde{Q} = 1$, whereas $\tilde{Q} = 0$ for maximally scattered distributions. Hence, the indeterminacy in the tangential component of v can be expected to increase with $\tilde{Q}$.

Closed-form solution. Assuming that A according to (20) is non-degenerate, the explicit solution is

$$v = -A^{-1}b \tag{24}$$

and the residual

$$r = c - b^T A^{-1} b. \tag{25}$$

For reasons to be explained in section 4.6, we also define the *normalized residual* as

$$\tilde{r} = \frac{r}{\operatorname{trace} A} = \frac{c - b^T A^{-1} b}{\operatorname{trace} A}. \tag{26}$$

If A is singular, or close to singular, it is preferable to use the pseudo inverse. In this 2-D case, it is given by

$$A^{\dagger} = \frac{1}{(\operatorname{trace} A)^2} A. \tag{27}$$

The pseudo inverse is preferred when the ratio between the singular values is sufficiently small, or equivalently the normalized anisotropy is sufficiently close to one.

In practice, the window function is chosen as a Gaussian kernel (with integration scale s), since then and only then the components of A satisfy scale-space properties under variations of s (which propagate to the distribution of gradient directions described by A as a composed object). Concerning the relation between s and the local scale t for computing derivatives, one should, in principle, consider a two-parameter variation. In the experiments to be presented, we have throughout used $s = \gamma^2 t$ with $\gamma = 2$.

4.3 Hierarchical and iterative flow field computations

By using scale-space operators at a certain scale t, it is, in general, only possible to capture disparities of the same order of magnitude as $\sqrt{t}$. This motivates a coarse-to-fine approach. Moreover, to reduce the approximation error in the local linearization, it is natural to compute iterative disparity updates, using the current disparity estimate $v^{(k)}$ when computing the brightness difference $R(\xi_L + v_L^{(k)}(\xi_L;\, t);\, t) - L(\xi_L;\, t))$, and iterating until R and L are in sufficient alignment.

If the transformation is not locally a pure translation, a higher order (*e.g.*, affine) model is required to reduce the approximation error, and corresponding compensations needed when computing the brightness differences. These iterations can be driven either by the affine scale-space and shape adaptation or by performing local warping and solving an extension of (19) with the locally constant flow model replaced by a local affine (see (Bergen *et al.* 1992; Barron *et al.* 1994; Lindeberg 1994b) for details).

4.4 Bidirectional matching and consistency measures

The previous matching scheme can be applied in both directions, which gives independent flow field estimates. A natural inconsistency measure is then

$$e_L(x_L;\, t) = v_L(x_L;\, t) + v_R(x_L + v_L(x_L;\, t);\, t),$$

and a natural measure of the strength of the response $R_L(x_L;\, t) = P_L(x_L;\, t)\, P_R(x_L + v_L(x_L;\, t);\, t)$, where P is the average square gradient magnitude. These entities and the normalized residual $\tilde{r}$ are then combined into the (heuristically chosen) confidence measure

$$W_L(x_L;\, t) = R_L(x_L;\, t)\, \exp(-\omega e_L^2/t)/(\tilde{r}_0 + \tilde{r}_L/t).$$

The motivations for this choice are that the significance should increase with the strength of the response and decrease with the inconsistency. The factors $1/t$ normalize the spatial errors with respect to the current scale, ω (here, ≈ 0.1) determines how large disparity inconsistencies are tolerated, and $\tilde{r}_0$ (here, ≈ 0.01) is a non-essential threshold to avoid divisions by zero.

4.5 Flow field correction and flow field smoothing

To suppress spurious errors, only disparity updates with $|v^{(k+1)}(x;\, t) - v^{(k)}(x;\, t)| < \nu\sqrt{t}$ propagate unaffected ($\nu \approx 2$). Larger updates are truncated.

Moreover, at each iteration, the flow field is smoothed using the confidence values W as weights

$$v'(x;\, t) = \frac{\int_{\xi \in \mathbb{R}^2} v(\xi;\, t)\, W(\xi;\, t)\, w_x(\xi;\, s(t))\, d\xi}{\int_{\xi \in \mathbb{R}^2} W(\xi;\, t)\, w_x(\xi;\, s(t))\, d\xi} \quad (28)$$

This leads to a rapid propagation of disparities from regions with strong variations to the interior of smooth regions. Moreover, spurious deviations are suppressed.

4.6 Scale selection

Within this framework, disparity estimates can be computed at any scale, using conceptually simple front-end operations. A fundamental problem, however, concerns how to combine the information from different scales. Selecting disparity estimates from the finest scales is not sufficient. These estimates can be very sensitive to noise and other interfering fine-scale structures. Unless explicit knowledge is available about what are the proper finest scales, this coarse-to-fine framework needs to be complemented by a mechanism for scale selection.

Intuitively, such a scale selection mechanism should select coarse-scale disparity fields from noisy data, for which fine-scale correspondences may be impossible to establish. Correspondingly, it should select fine-scale representatives from the disparity fields from sharp data that contain detailed information, so as to produce a maximally accurate disparity field.

Selection method. Clearly, the residual (25) depends upon the local contrast and cannot be used for such judgements. A straightforward but nevertheless powerful approach is to *select the scale that minimizes the normalized residual (26) over scales.* A basic motivation for the specific definition (26) is that the division cancels the effect the local brightness variations and trace A is a natural measure of the strength of the response. Since the dimensions involved are as follows:

Entity	Dimension
A	[luminance]^2/[length]^2
b	[luminance]^2/[length]
c	[luminance]^2

the normalized residual has dimension [length]^2 and reflects a spatial error in the disparity estimate.

Qualitative effects. Relating to the abovementioned intuitive requirements, the qualitative effects of this scale selection method are as follows:

At too coarse scales, a uniform deformation model cannot be expected to hold over the entire region. Also, the shape distortions can be expected to be stronger, thereby increasing the normalized residual.

At too fine scales, where noise and other fine-scale structures are present, the likelihood that these structures obey the same motion model will be low. Hence, the normalized residual can be expected to increase.

Selecting the minimum leads to a natural trade-off between these effects.

4.7 Experiments

Figure 1 shows the result of applying the composed scheme to synthetic patterns transformed by a pure expansion and a pure rotation, respectively. 10% white Gaussian noise added to each image after the transformation. Notice, how well the flow fields are captured. A numerical evaluation shows that the accuracy in the estimates corresponds to sub-pixel accuracy. For a more extensive evaluation, see (Lindeberg 1994b).

Figure 2 shows corresponding results for a detail of a head subject to a rather large (unknown) rotation. Note that except for the upper right corner, where most points either correspond to occluded points or points outside the image, a correct matching has been obtained without any use of epipolar geometry.

4.8 Summary and discussion

We have considered the problem of estimating image deformations using visual front-end operations, *i.e.* scale-space smoothing, derivative computations and pointwise combinations of these primitives. The framework builds upon schemes for computing optic

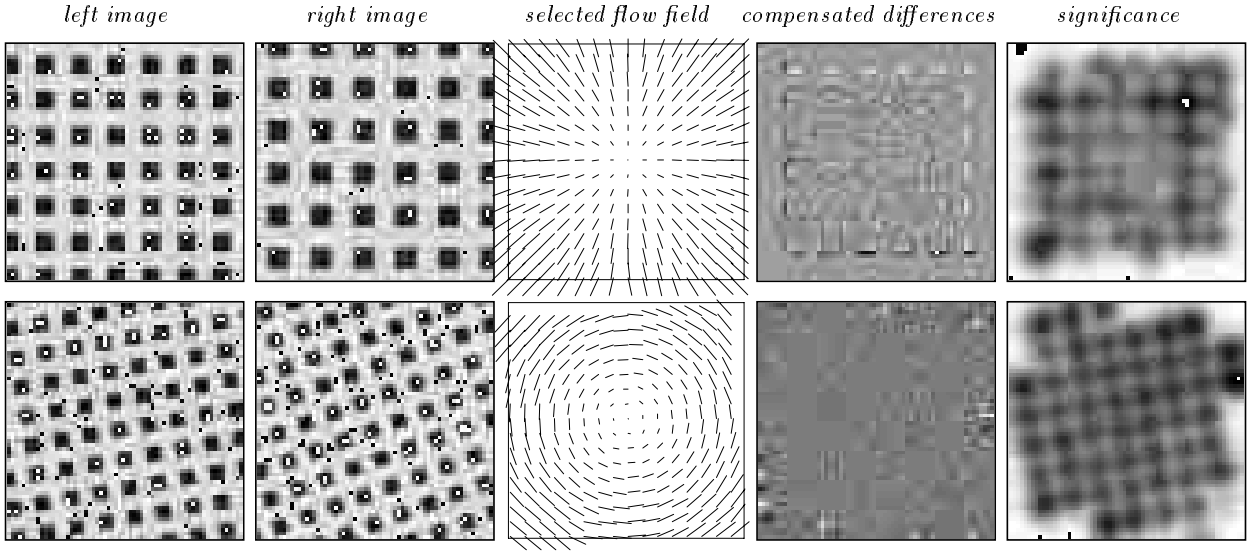

Figure 1: Flow fields computed using the proposed scheme with automatic scale selection: (top row) synthetic expansion with 10% noise, (bottom row) synthetic rotation with 10% noise. The columns show from left to right; (a) left image, (b) right image, (c) estimated flow field (left), (d) compensated differences, (e) significance measure. (Image size: 64×64.)

flow with explicit mechanisms for hierarchical and iterative updating, bidirectional matching, and confidence measurements. In addition to these components, a method has been included for selecting the scales at which the deformation estimates should be extracted. This method is based upon minimizing a normalized residual over scales and has the intuitively appealing property of selecting coarser scale estimates in the presence of noise and locally inconsistent estimates.

An interesting aspect of the resulting approach is that the computed information is contained in the control signals for bringing the image data into alignment.

In an active situation, these signals can serve as a natural vergence mechanism. If the translation based scheme is applied in the log-polar domain, it provides a lower order approach for measuring the other primitive transformations in the similarity group, *i.e.*, rotations and uniform size changes. When extended to local full affine models, the scheme allows for unbiased estimation of the invariant first-order flow components.

5 Enforcing consistency

For deformation estimates that have been computed independently, it is not guaranteed that shape descriptors derived from them correspond to a coherent surface; *e.g.*, for a field of surface orientations to correspond to a depth map, the rotation must be zero.

Figure 3 shows an example of enforcing such consistency on monocular data by fitting a pointwise (and hence parameter free) depth map to a field of surface orientation estimates computed by a slight modification of the shape from texture method in (Lindeberg and Gårding 1993). For each point, the surface orientation estimate has been obtained from a *centered second moment matrix*

$$\nu = \int_{\xi \in \mathbb{R}^2} (\nabla L)(\nabla L)^T \, w \, d\xi - (\overline{\nabla L})(\overline{\nabla L})^T \quad (29)$$

where $\overline{\nabla L} = \int_{\xi \in \mathbb{R}^2} \nabla L \, w \, d\xi$ and w is a Gaussian window function. (This descriptor obeys a similar linear transformation property $\nu_L(q) = A^T \, \nu_R(p) \, A$ as the non-centered second moment matrix μ. The major differences are that it is invariant to superimposed linear gradients $L \mapsto L + c_1 + c_2^T x$ and less sensitive to small perturbations of the centers of blob-like surface structures.) From (a modification of) the weak isotropy assumption—that ν in the surface should be a constant times the unit matrix—the slant angle has been computed as $\arccos(\sigma_2/\sigma_1)$ and the tilt direction from $\psi/2$. (For more details about the algorithm, see (Lindeberg 1994b)). Observe how the qualitative shape of the torso is captured by these very simple operations.

References

L. Alvarez, F. Guichard, P.-L. Lions, and J.-M. Morel. Axioms and fundamental equations of image processing. *Arch. for Rational Mechanics*, 123(3):199–257, 1993.

J. J. Barron, D. J. Fleet, and S. S. Beachemin. Performance of optical flow techniques. *IJCV*, 12(1), 1994.

J. Bergen, P. Anandan, K. Hanna, R. Hingorani. Hierarchical model-based motion estimation. *2nd ECCV*, 237–252, 1992.

L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever. Scale and the differential structure of images. *IVC*, 10(6):376–388, 1992.

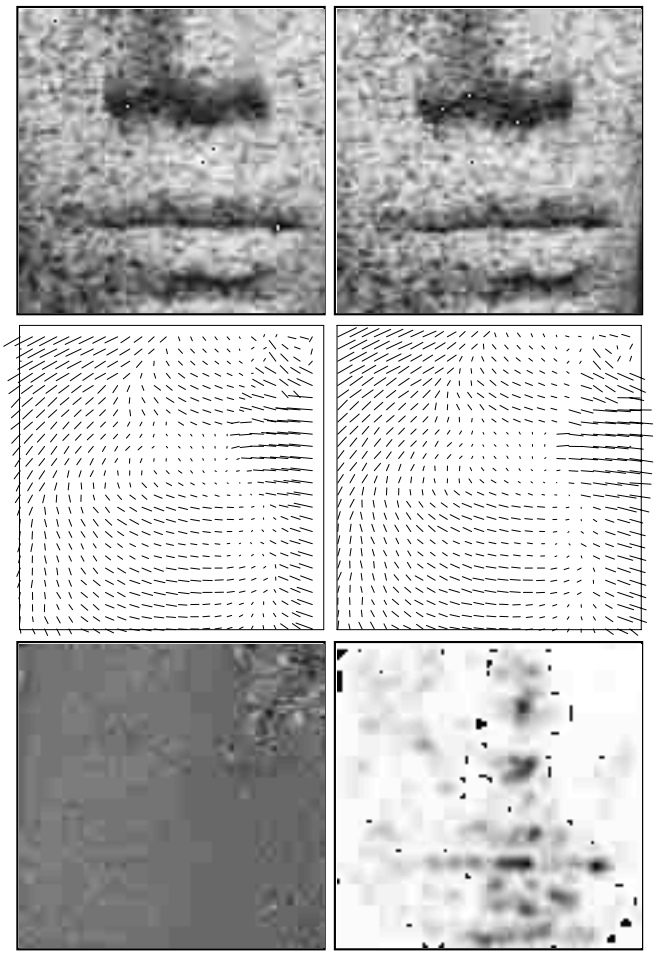

Figure 2: Flow fields computed from a detail of a statue. From top left to bottom right: (a)–(b) left and right images, (c)–(d) left and right flow fields, (e) left compensated differences, (f) left significance. (Image size: 200×200.)

B. K. P. Horn and B. G. Schunck. Determining optical flow. *AI*, 17:185–204, 1981.

D. G. Jones and J. Malik. A computational framework for determining stereo correspondences from a set of linear spatial filters. *2nd ECCV*, 395–410, 1992.

J. J. Koenderink and A. J. van Doorn. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta*, 22(9):773–791, 1975.

J. J. Koenderink and A. J. van Doorn. Receptive field families. *Biol. Cyb.*, 63:291–298, 1990.

T. Lindeberg and J. Gårding. Shape from texture from a multi-scale perspective. *4th ICCV*, 683–691, 1993.

T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Acad. Publ., Dordrecht, Netherlands, 1994a.

T. Lindeberg. Direct estimation of affine deformations of brightness patterns using visual front-end operations with automatic scale selection. ISRN KTH/NA/P–94/33–SE, 1994b.

R. Manmatha. Measuring the affine transform using Gaussian filters. *3rd ECCV*, vol. 801, 159–164, 1994.

M. Proesmans, L. van Gool, E. Pauwels, and A. Oosterlinck. Determination of optical flow and its discontinuities using non-linear diffusion. *3rd ECCV*, vol. 801, 295–304, 1994.

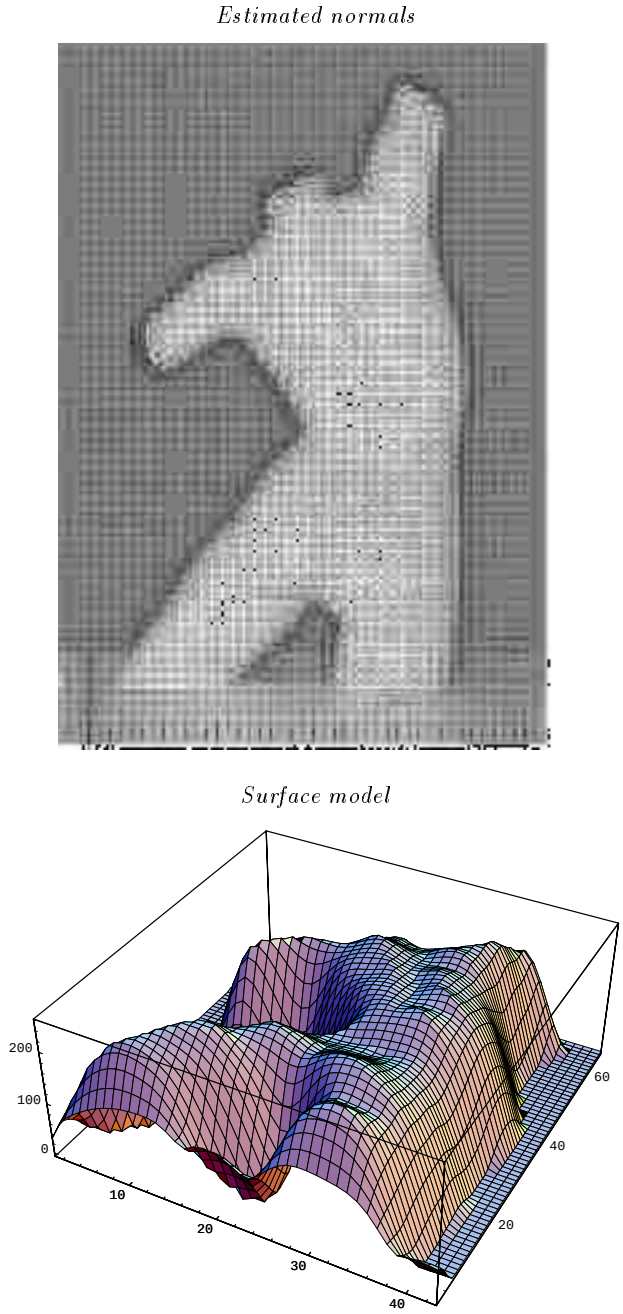

Figure 3: (top) Surface normals estimated from affine deformations measured by centered second moment matrices. (bottom) Surface model constructed by enforcing internal consistency (zero rotation) on the field of surface normals by least squares fitting of a pointwise depth map.

G. Sapiro and A. Tannenbaum. Affine invariant scale-space. *IJCV*, 11(1):25–44, 1993.

J. Sato and R. Cipolla. Extracting the affine transformation from texture moments. *3rd ECCV*, vol. 801, 165–172, 1994.

P. Werkhoven and J. J. Koenderink. Extraction of motion parallax structure in the visual system. *Biol. Cyb.*, 1990.

A. P. Witkin. Scale-space filtering. *8th IJCAI*, 1019–1022, 1983.