THE GENERALIZATION, OPTIMIZATION, AND INFORMATION-THEORETIC JUSTIFICATION OF FILTER-BASED AND AUTOCOVARIANCE-BASED MOTION ESTIMATION.

Rudolf Mester

Institute for Applied Physics, Image and Vision Group
Robert-Mayer-Str. 2–4, D–60054 Frankfurt am Main, Germany
mester@iap.uni-frankfurt.de www.uni-frankfurt.de/fb13/iap/cvg

ABSTRACT

We discuss the theoretical foundations of measuring motion in video data, and relate this strongly to statistical estimation theory. A very general class of motion estimation methods is characterized by determining second order moments of filter bank outputs. These moments are represented in tensors, and motion estimation boils down to analyzing their eigensystems. An alternative approach is to directly estimate and analyze the autocorrelation of the given signal. We provide motivation for developing these approaches further towards directional entropy rate criteria rather than rely on conventional directional smoothness criteria. This paper emphasizes that prior knowledge on the video signal (e.g. spatial autocovariance, distribution of expected motion speed, noise spectrum, ...) should be integrated into the motion estimation procedure. Relations between different classes of motion algorithms (differential, tensor-based, steerable filters ...) are discussed and perspectives for a unification and enhancement of such procedures are presented.

1. INTRODUCTION

After decades, motion estimation for video data still provides a variety of challenges, most notably in terms of robustness, precision and computational effort. The diversity of proposed analysis methods increases the problem of selecting and optimizing a motion estimation procedure, while relations and equivalence between different approaches are often only scarcely visible.

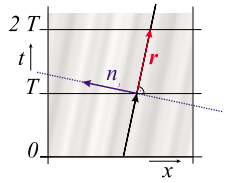
In order to achieve further improvements with respect to established motion estimators, it is useful to exploit explicit knowledge about the spatio-temporal statistics of the video signal and the noise, about the acquisition process, and the statistics of the motion itself. On the basis of realistic models for these entities, the relations between apparently different classes of motion estimation algorithms can be uncovered further.

2. MODELS FOR ORIENTED SIGNALS AND SPATIO-TEMPORAL FLOW

Let us regard a three-dimensional space-time volume spanned by the coordinates $(x,y,t) \equiv (x_1,x_2,x_3) \equiv \mathbf{x}$, $\mathbf{x} \in \mathbb{R}^3$ on which the signal $s(\mathbf{x})$ is defined. A local neighborhood with *ideal orientation* can be described as

$$s(\mathbf{x}) = f_2(\ \mathbf{n}_1^T \cdot \mathbf{x},\quad \mathbf{n}_2^T \cdot \mathbf{x}\ )\ , \tag{1}$$

where $f_2(\cdot,\cdot)$ is a scalar function of two arguments, and $\mathbf{n}_1, \mathbf{n}_2$ are two orthonormal vectors perpendicular to the local motion vector $\mathbf{r}$. The following figure illustrates this by an x,t cross-section through a space-time volume (x,y,t) for the case of a signal moving with constant speed.



We denote each signal $s(\mathbf{x})$ for which it is possible to find a 3D coordinate frame $\mathbf{n}_1, \mathbf{n}_2, \mathbf{r}$ according to eq.1 (i.e. such that the signal varies only in 2 directions), as a *rank 2 signal in a three-dimensional space* [4].

3. DIFFERENTIAL APPROACHES TO MOTION ANALYSIS

The general principle behind all differential approaches is that the rank condition is understood as the conservation of some local image characteristic throughout its temporal evolution; this is reflected in terms of differential-geometric descriptors. In its simplest form, the assumed conservation of brightness along the motion trajectory through space–time leads to the well-known *brightness constancy constraint equation* (BCCE), where $\mathbf{g}(\mathbf{x})$ is the gradient of the

gray value signal $s(\mathbf{x})$:

$$\left(\frac{\partial s}{\partial x_1}, \frac{\partial s}{\partial x_2}, \frac{\partial s}{\partial x_3}\right) \cdot \mathbf{r} = 0 \quad \Leftrightarrow \quad \mathbf{g}^T(\mathbf{x}) \cdot \mathbf{r} = 0. \qquad (2)$$

Since $\mathbf{g}^T(\mathbf{x}) \cdot \mathbf{r}$ is proportional to the directional derivative of s in direction $\mathbf{r}$, the BCCE states that this derivative vanishes in the direction of motion. Of course, one equation of type $\mathbf{g}^T(\mathbf{x}) \cdot \mathbf{r} = 0$ is not enough for determining the sought vector $\mathbf{r}$ (which has two degrees of freedom). Besides that, real image signals are never true rank-2 signals. By spatial averaging using a weight mask $w(\mathbf{x})$, we obtain

$$\int_V w(\mathbf{x}) \cdot \left|\mathbf{g}^T(\mathbf{x}) \cdot \mathbf{r}\right|^2 d\mathbf{x} \longrightarrow \min$$

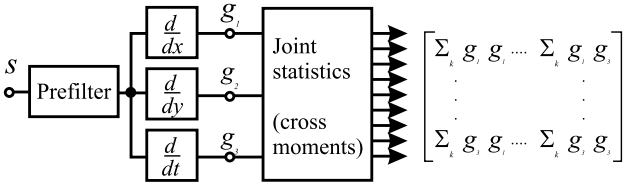
$$\Rightarrow \quad \int_V \mathbf{r}^T \cdot \left(\mathbf{g}(\mathbf{x}) \cdot w(\mathbf{x}) \cdot \mathbf{g}^T(\mathbf{x})\right) \cdot \mathbf{r} \; d\mathbf{x} \longrightarrow \min$$

$$\Rightarrow \quad \mathbf{r}^T \cdot \mathbf{C}_g \cdot \mathbf{r} \longrightarrow \min \qquad (3)$$

$$\text{with} \quad \mathbf{C}_g := \int_V \mathbf{g}(\mathbf{x}) \cdot w(\mathbf{x}) \cdot \mathbf{g}^T(\mathbf{x}) \, d\mathbf{x} \qquad (4)$$

Given the variability (in terms of autocovariance) of the quasi-constant target entity $\mathbf{r}$, application of weighted least squares theory determines the optimum weighing function $w(\mathbf{x})$, very similar to *normalized convolution* [8]. The solution vector $\mathbf{r}$ is the eigenvector corresponding to the minimum eigenvalue of the *structure tensor* $\mathbf{C}_g$ [1, 5, 6, 7]. The minimization criterion according to eq.3 can be replaced by modified criteria which yield exactly the same solution $\mathbf{r}$ in case of ideal rank-2 signals, but different solutions for the realistic case of perturbations in matrix $\mathbf{C}_g$, see [10, 11].

The definition of the 'classical' structure tensor $\mathbf{C}_g$ shows that $\mathbf{C}_g$ encapsulates the 2nd order joint statistics of the gradient. More precisely, $\mathbf{C}_g$ comprises the *estimates* of the second order cross moments of the components of the *estimated* gradient.



Estimation of the spatio-temporal gradient is a formidable problem in itself since only discrete samples of the underlying signal are available. In most cases, a *prefilter* 'regularizes' the gradient computation; it should be designed according to maximizing the S/N ratio in the output entities [9].

3.1. Relation between differential and acf-based approaches

Let $\hat{\mathbf{g}}(\mathbf{x})$ be an estimate of the gray value gradient at position $\mathbf{x}$ on the image lattice. The components $\hat{g}_1$, $\hat{g}_2$ and $\hat{g}_3$ of $\hat{\mathbf{g}}$ are computed by linear convolution filters and therefore given by linear expressions of type

$$\hat{g}_u(\mathbf{x}) = \sum_{i=-n}^{n} \sum_{j=-m}^{m} \sum_{k=-p}^{p} \alpha_{uijk} \cdot s(\mathbf{x} - \mathbf{d}_{ijk}) \qquad u \in \{1,2,3\} \, .$$

Therefore, the cross moments $\hat{c}_{guv}$, $u, v \in \{1,2,3\}$ that form the entries of the estimated structure tensor $\hat{\mathbf{C}}_g$ can be written as a weighted linear combination of product terms between signal samples:

$$\hat{c}_{guv} = \sum_{\mathbf{x},\mathbf{d}} \beta(\mathbf{d}) \cdot s(\mathbf{x}) \cdot s(\mathbf{x} - \mathbf{d})$$

From this, we see that the $\hat{c}_{guv}$ can be derived directly from the cross moments of $s(\mathbf{x})$ and the corresponding values $s(\mathbf{x} - \mathbf{d})$ at a large, but finite set of given displacements $\mathbf{d} = (d_1, d_2, d_3)$. In other words, the initially sought entities (cross moments of the gradient components) are directly and linearly related to the estimate of the three–dimensional autocorrelation function $\varphi_{ss}(d_1, d_2, d_3)$. This is a clear clue that motion estimation could be based directly on the autocorrelation function.

3.2. Generalizing differential approaches

The differential formulation of brightness constancy along the motion trajectory is not the unique and presumably not the most expressive way of specifying a relation between the entity that is sought (the motion vector $\mathbf{r}$) and the data that can be observed. As it is well known, the BCCE describes the situation for a *continuous signal*, and it does not explicitly consider the different error terms that are caused by observation noise, spatio-temporal pixel aperture, and by the necessary discretization of the problem.

Beyond that, the formulation in terms of derivatives or gradients does not lend itself so much for the development of motion estimation procedures that take into account the *spectral characteristics of the image signal* and the *spectral characteristics of the noise*. Quite obviously, since any pure rank-2 signal s may be prefiltered by any filter with radially symmetric transfer function $P(\mathbf{f})$ without changing the *direction* of the eigenvectors of $\mathbf{C}_g$, the interpretation of $\mathbf{g}$ as being the local gradient is much too narrow.

Assuming brightness constancy along the motion trajectory, all higher order directional derivatives vanish in the motion direction:

$$\frac{\partial s}{\partial \mathbf{r}} \stackrel{!}{=} 0 \quad \bigcap \quad \frac{\partial^2 s}{\partial \mathbf{r}^2} \stackrel{!}{=} 0 \quad \bigcap \quad \ldots \qquad (5)$$

A condition which is less stringent than eq. 5 can be obtained by summing up the constraints:

$$\alpha_1 \frac{\partial s}{\partial \mathbf{r}} + \alpha_2 \frac{\partial^2 s}{\partial \mathbf{r}^2} + \alpha_3 \frac{\partial^3 s}{\partial \mathbf{r}^3} \stackrel{!}{=} 0 \qquad (6)$$

The left hand side of this equation is nothing else than a generator for a very rich class of filter operators, parameterized by the direction vector $\mathbf{r}$:

$$h(\mathbf{x} \mid \mathbf{r}) * s(\mathbf{x}) \stackrel{!}{=} 0$$

where $*$ denotes the convolution operator.

Like in the case of the normal BCCE, this equation will be satisfied almost never. Thus, we end up with optimization criteria like

$$\int_{\mathbf{x}} w(\mathbf{x}) \cdot |h(\mathbf{x} \mid \mathbf{r}) * s(\mathbf{x})|^2 \, d\mathbf{x} \quad \longrightarrow \min \tag{7}$$

where $h(\mathbf{x} \mid \mathbf{r})$ comprises the combination of directional derivatives of different order, and an optional *pre-filter* $p(\mathbf{x})$.

This means: Eq.7 minimizes the frequency-weighted directional variation of the signal in the direction of motion. The weight functions (now in the spectral domain!) implicitly contained in the layout of the operator $h(\mathbf{x} \mid \mathbf{r})$ have to be designed according to the power spectrum of signal and noise, or (equivalently) according to their covariance structure. Furthermore, the range of expected motion does strongly influence the covariance structure and should be considered (see [9]). This all gives us a direct hint towards the general theory of *linear prediction* (see section 5) and offers ways to integrated knowledge on the statistical structure of signal and noise.

4. PROJECTION CRITERIA FOR MOTION ESTIMATION

In this section, we will come back to the model for oriented space/time signals sketched in section 2. Let $s = s(\mathbf{x})$ be the observed signal. The main idea is to subdivide the given signal in two terms according to $s = \hat{q} + u$, where $\hat{q}$ is an estimate of the 'true' signal q which is supposed to be a sample from the class C_o of rank–2 oriented signals, and u is a residual signal. Under the condition that all realizations q_i from the class C_o of oriented signals are equally likely, a probabilistic approach to motion estimation would be to find the specific pair $\{\hat{q}_i, u_i\}$ that fulfills the condition $\hat{q}_i + u_i = s$ and which maximizes the likelihood of the noise realization u_i. If the noise is modeled as a zero-mean quasi-white[1] Gaussian process, this leads directly to a least squares fitting procedure. If additionally the class C_o in itself is also modelled probabilistically, a Bayesian method will be obtained which boils down to the optimization problem

$$\text{JointProb}\,[\hat{q}_i, u_i]|_{\hat{q}_i + u_i = s} \longrightarrow \max\,, \tag{8}$$

but only the simpler case of ML estimation is discussed here. We start with minimizing the weighted energy of the residual u

$$\int_V w(\mathbf{x}) \cdot u^2(\mathbf{x}) \, d\mathbf{x} = \int_V w(\mathbf{x}) \cdot (s(\mathbf{x}) - \hat{q}(\mathbf{x} \mid \mathbf{r}))^2 \, d\mathbf{x} \longrightarrow \min$$

Here, $\hat{q}(\mathbf{x} \mid \mathbf{r})$ denotes an estimate of q, i.e. a projection of the given signal $s(\mathbf{x})$ on the (sub)space of signals oriented in direction $\mathbf{r}$. Under the assumption of brightness constancy in $\mathbf{r}$ and quasi-white noise, it also comprises a projection[2] in a quite different sense: let R be a ray in direction $\mathbf{r}$ through the regarded space-time volume V. Then the maximum likelihood estimate for the true gray value along this ray is

$$\begin{aligned} \hat{m} &= \frac{1}{L} \int_R w(\mathbf{x}) \cdot s(\mathbf{x}) \, d\mathbf{x} \\ \text{where} \quad L &:= \int_R 1 \cdot w(\mathbf{x}) \, d\mathbf{x}\,. \end{aligned}$$

With this estimate for the constant gray value $m(R \mid \mathbf{r})$ along a ray R in direction $\mathbf{r}$, the energy of the residual signal $s(\mathbf{x}) - \hat{m}$ yields a quadratic error measure

$$J_1(R) := \int_R w(\mathbf{x}) \cdot (s(\mathbf{x}) - \hat{m})^2 \, d\mathbf{x}\,. \tag{9}$$

A more general formulation that considers a frequency-dependent power spectrum of the noise is

$$J_2(R) := \int_R w(\mathbf{x}) \cdot [b_2(\mathbf{x}) * (s(\mathbf{x}) - \hat{m})]^2 \, d\mathbf{x}\,, \tag{10}$$

where $b_2(.)$ is the impulse response of a suitably chosen filter. Both for J_1 or J_2, the value of such a criterion has to be integrated over the 'remaining' directions of the regarded space/time volume V, and the direction $\mathbf{r}$ yielding the minimum value of the result is the sought motion direction.

Up to now, we have assumed that the transversal profile of the rank-2 signal $q(\mathbf{x})$ is arbitrary. Since this will not be a realistic assumption in most cases, it appears to be a natural step to restrict these signal variations to be 'smooth' in some sense. The criteria considered before will have to be modified again, this time with the effect of using a three-dimensional operator $b_3(\mathbf{x} \mid \mathbf{r})$ which is parameterized by $\mathbf{r}$.

$$J_3(\mathbf{r}) := \int_V w(\mathbf{x}) \cdot [b_3(\mathbf{x} \mid \mathbf{r}) * (s(\mathbf{x}) - \hat{q}(\mathbf{x}))]^2 \, d\mathbf{x}\,, \tag{11}$$

This is basically nothing else than performing an optimal LS (Wiener) filter restoration of the rank-2 signal, with direction vector $\mathbf{r}$ as the parameter controlling the filter, and the application of a suitable metric on the residual signal

between the observed signal s and the estimate $\hat{q}$. A even more general approach, allowing for higher order nonlinearities is:

$$J_4(R) := \mathsf{B}\,(s(\mathbf{x}) - \hat{q}(\mathbf{x}) \mid \mathbf{r}) \tag{12}$$

[1] = wide band noise with a flat isotropic power spectrum

[2] Of course, this projection approach has a close relation to the Radon transform which is not discussed here.

Here, B is a functional on the set of scalar functions defined on V. If this functional B (or the operator b_3 in eq. 11) is invariant with respect tothe directional mean value $\hat{m}$,then the computation of $\hat{m}$ is obviously not necessary. This is the case, for instance, for functionals that are based on one (or several) derivative kernel(s), like in the example of eq. 3. It is not very astonishing that most established methods,including steerable and directional filters, can be subsumed under this class of approaches. The specific functional $\int_V \left|(\nabla s)^T \cdot \mathbf{r}\right|^2 d\mathbf{x}$ discussed in section 3 has the advantage that it can be implemented as a steerable filter [13] parameterized by $\mathbf{r}$.

5. INFORMATION-THEORETIC JUSTIFICATION

Up to now, we have expressed the constancy (or small variation) of the signal along the motion direction by differential criteria. Alternatively, we might consider the entropy rate $\mathcal{H}$ of a (one-dimensional) stochastic process; it quantifies the average flux of information (as measured in Shannon's sense) as provided by a the process ([2],p.274). It is a monotonic function of the minimum mean squared error which an optimum predictor of a sample of the process would yield, given the infinite past. Since the entropy rate $\mathcal{H}$ is zero if and only if the function is constant, the entropy rate of an ideal translational motion signal is zero along the direction of motion. In contrast to conventional differential geometric measures (BCCE and generalizations), it allows for certain variations of the moving objects. It is merely necessary to formulate explicit models of the variations to expect, such as slow illumination changes. By doing so, models extensively used in the evaluation of video coding schemes (but much less in the analysis of motion) can be exploited, for instance the theory of motion compensated prediction [3].

Of course, the design of the optimum (linear) predictor and the minimum value of the prediction error variance are again related to the autocovariance structure of the image material. Therefore, it is required to determine the autocovariance function (ACF) at all discrete grid points in a small subvolume of the x,y,t volume. The dimensions to be selected for this volume depend on the average ACF structure of the signal, to be determined *a priori* from the 2D image ACF and the range of expected motion (see [9] for details).

If the *spatial* correlation (i.e. in the x,y-plane) is compensated for by a statically designed whitening or innovations filter ([12], p.402), the predictor operates only in the hypothetic motion direction; it is then a conventional one-dimensional predictor that depends on the values of the measures ACF. Let $\varphi_{ss}(n)$ be an ordered discrete ensemble of ACF values in the regarded direction. It is not necessary to explicitly design the predictor, the only entity that in fact has to be computed is the minimum prediction error variance which can be shown to be of the form

$$Q_{min} = \varphi_{ss}(0) - \mathbf{p}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{p} \tag{13}$$

where the elements of $\mathbf{p}$ are the different $\varphi_{ss}(n), n = 1,N$, and $\mathbf{C}$ is the covariance matrix which is also built from the discrete ACF values $\varphi_{ss}(n)$ (cf. [12]). Experimental evaluation of this approach is currently ongoing.

6. CONCLUSIONS

This paper contributes to the theoretical foundations of motion estimation and strengthens its relations to statistical signal processing. It extends conventional schemes such as the classical BCCE based motion estimation algorithms and relates these differential approaches to the alternative of analyzing the autocovariance structure of image signals. It shows ways to introduce explicit models for the statistical image structure and provides a new information-theoretic view on motion analysis.

7. REFERENCES

[1] J. Bigun and G. H. Granlund. Optimal orientation detection of linear symmetry. In *First International Conference on Computer Vision, ICCV (London)*, pages 433–438, Washington, DC., June 8–11 1987. IEEE Computer Society Press.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1st edition, 1991.

[3] B. Girod. Motion-compensating prediction with fractional pel accuracy. *IEEE Transactions on Communications*, 41(4), April 1993.

[4] G. H. Granlund and H. Knutsson. *Signal processing for computer vision*. Kluwer, 1995.

[5] H. Haussecker and H. Spies. Motion. In *Handbook of Computer Vision and Applications*. Academic Press, 1999.

[6] B. Jähne. *Digital Image Processing*. Springer Verlag, 4th edition, 1998.

[7] B. Johansson and G. Farnebäck. A theoretical comparison of different orientation tensors. In *Proceedings SSAB02 Symposium on Image Analysis*, pages 69–73, Lund, March 2002.

[8] H. Knutsson and C.-F. Westin. Normalized and differential convolution: Methods for interpolation and filtering of incomplete and uncertain data. June 1993.

[9] R. Mester. A generalization of differential optical flow estimation using 3d covariance functions of signal and noise. In *Submitted in April 2003*.

[10] M. Mühlich and R. Mester. Subspace methods and equilibration in computer vision*. Technical Report XP-TR-C-21, Frankfurt University, 1999.

[11] O. Nestares, D. J. Fleet, and D. J. Heeger. Likelihood functions and confidence bounds for Total Least Squares estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'2000)*, Hilton Head, 2000.

[12] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition, 1991.

[13] E. P. Simoncelli. Design of multi-dimensional derivatives filters. In *Intern. Conf. on Image Proc.*, Austin TX, 1994.