Optimal Schemes for Motion Estimation on Colour Image Sequences

Julian Magarey
jfam@cssip.edu.au
CSSIP
SPRI Building
The Levels, SA 5095
Australia

Anil Kokaram
ack@eng.cam.ac.uk

Nick Kingsbury
ngk@eng.cam.ac.uk

Signal Processing Laboratory
Cambridge University Engineering Department
Trumpington St, Cambridge CB2 1PZ
United Kingdom

Abstract

This paper describes a method for incorporating the chrominance information when estimating motion in a colour image sequence. It is based on a Maximum Likelihood formulation of the motion estimation problem which assumes homogeneous additive Gaussian noise in each colour component, with known inter-field correlation statistics. The formulation is applied to the complex-wavelet-domain matching algorithm of Magarey and Kingsbury [1]. We also define a noise-decorrelating colour space transform which provides a simple implementation of the ML formulation in the wavelet domain. Results for noisy synthesised colour sequences with known motion and noise statistics demonstrate the superiority of the exact ML formulation over straightforward, unweighted three-component estimation, most noticeably in high noise conditions.

1 Introduction

Motion estimation (ME) for image sequence manipulation has been crucial to the development and successful deployment of many video processing tasks, e.g. video coding, video restoration, and computer vision. Traditionally, extraction of the motion information has been limited to the use of the luminance or intensity information in the scene. The exclusion of chrominance information has been forced by the need to limit the complexity of the algorithms and the observation that many of the visual cues for motion estimation can indeed be found in the luminance information. However, it is readily acknowledged that chrominance information must improve the performance of motion estimation algorithms if only through the ability to take advantage of the entire data set available. Work on the use of colour for motion estimation [2, 3, 4] has extended the monochrome maximum likelihood (ML) formulation of motion estimation to deal with vector valued (colour) frames. Although these works identify the improvement gained with the use of colour, they assume that each colour component contributes equally to the final solution. This is based in turn on the implicit assumption that the model noise is uncorrelated and equivariant between the three colour components or *fields* which make up each frame.

However, this is often far from being the case. For example, if the component noise is uncorrelated in RGB-space it is correlated in YUV-space. This paper addresses the need for a careful analysis of the optimal incorporation of colour information in motion estimation. We follow the approach of Konrad and Dubois [3] in extending the ML motion estimator to deal with vector inputs. In previous work [5], we applied the generalised ML formulation to two standard ME algorithms which operate in the original pixel domain: region-based matching and gradient-based. In Section 2 we apply it to the complex-wavelet-domain monochrome ME algorithm of Magarey and Kingsbury [1, 6]. It is shown that a noise-decorrelating colour-space transform, when applied to the original components, reduces the problem to the simple case of independent and equal contributions from each component, significantly reducing the amount of computation required. Performance comparisons are based on the deviation of the estimated motion fields from the known motion fields in synthesised sequences. The results in Section 3 show that the true vector ML formulation outperforms straightforward unweighted RGB-space estimation in terms of robustness to correlated additive noise. Our results also suggest that the most efficient strategy would be adaptive, based primarily on luminance and only incorporating chrominance appropriately where required. This is confirmed by some results obtained from a real image sequence.

2 Wavelet domain MLME

2.1 Sequence modelling

Let us consider the ML formulation of the multicomponent ME problem. To this end, we begin with the common assumption of *intensity conservation,* which requires that intensity *in each component* is constant along the motion

0-8186-8183-7/97 $10.00 © 1997 IEEE

trajectory defined by the displacement $\hat{\mathbf{d}}(\mathbf{x})$ at pixel $\mathbf{x}$. This model is corrupted by a 3-component vector $\mathbf{e}_n(\mathbf{x})$:

$$\mathbf{g}_n(\mathbf{x}) = \mathbf{g}_{n-1}(\mathbf{x} + \hat{\mathbf{d}}(\mathbf{x})) + \mathbf{e}_n(\mathbf{x}) \qquad (1)$$

The quantity $\mathbf{e}_n(\mathbf{x})$, which is primarily due to observation noise, is modelled as zero-mean Gaussian, with (3-by-3) covariance matrix $R_\mathbf{g}(\mathbf{x})$. Its probability density function (pdf) is therefore

$$p(\mathbf{e}_n(\mathbf{x})) \propto \exp\left(-\frac{1}{2}\mathbf{e}_n^T(\mathbf{x}) R_\mathbf{g}(\mathbf{x})^{-1}\mathbf{e}_n(\mathbf{x})\right) \qquad (2)$$

The direct ML formulation estimates the translation model parameter $\hat{\mathbf{d}}$ as the maximising argument of the pdf of $\mathbf{g}_n(\mathbf{x})$ given $\mathbf{g}_{n-1}(\mathbf{x})$ and (2), with its parameter $R_\mathbf{g}$.

In a previous paper [5] we showed how to obtain a robust ML estimate of $\hat{\mathbf{d}}$ over a region Ω of pixels $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ centred on $\mathbf{x}$ (this is the common assumption of *constant local flow* for regularising ME algorithms). The other assumptions were that model deviations are homogeneous over the image (allowing us to drop the $\mathbf{x}$ argument from $R_\mathbf{g}$), and the absence of noise correlation between different pixels in the region. The result is the region-based ML matching algorithm for vector images. A Taylor expansion of each component, involving intensity gradients, may be used to find a closed-form approximate solution by linear least-squares. This is the *gradient-based* vector ML algorithm [5].

2.2 The CDWT decomposition

The vector ML formulation may be applied to the complex wavelet domain matching algorithm of Magarey and Kingsbury [1]. This algorithm is based on a linear transform, dubbed the Complex Discrete Wavelet Transform (CDWT), which is implemented by repeatedly applying a separable filter-downsample building block comprising a Gabor-like basis pair of 4-tap, rational-valued complex filters. The result is an efficient structure which decomposes each component g_k into a multiresolution pyramid of oriented complex subimages $\{g_k^{(n,m)}, m = 1, \ldots, m_{max}, n = 1, \ldots, 6\}$ (m indexes scale, while each n corresponds to a specific orientation). Each subimage may be regarded as the output of a 2-d (separable) Gabor (Gaussian windowed bandpass) filter, downsampled by 2^m in each direction. The vector subimages of frame 1 are $\mathbf{g}_1^{(n,m)} = [g_{1,1}^{(n,m)} g_{1,2}^{(n,m)} g_{1,3}^{(n,m)}]^T$.

2.3 CDWT domain matching

The CDWT algorithm performs efficient matching between corresponding subpixels in the complex subimages. Starting at the level of coarsest resolution ($m = m_{max}$), we postulate a uniform translation $\hat{\mathbf{d}}(\mathbf{x})$ over all the pixels in the support region of each subpixel $\mathbf{x}$ (i.e. the implicit regions of constant local flow are defined by the corresponding Gabor filter). Because the CDWT uses linear filtering, we can rewrite (1) for subpixel $\mathbf{x}$ of subimage (n,m) of frame 2 as

$$\mathbf{g}_1^{(n,m)}(\mathbf{x} + \hat{\mathbf{f}}(\mathbf{x})) - \mathbf{g}_2^{(n,m)}(\mathbf{x}) = \mathbf{e}^{(n,m)}(\mathbf{x}) \qquad (3)$$

where $2^m\hat{\mathbf{f}}(\mathbf{x}) = \hat{\mathbf{d}}(\mathbf{x})$ and $\mathbf{e}^{(n,m)}$ is the model error in subband (n,m), which by the assumption of homogeneity is independent of $\mathbf{x}$. The joint pdf of $\mathbf{e}^{(n,m)}$ is

$$p(\mathbf{e}^{(n,m)}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{e}^{(n,m)}\right)^H (R_\mathbf{g}^{(n,m)})^{-1}\mathbf{e}^{(n,m)}\right) \qquad (4)$$

where $R_\mathbf{g}^{(n,m)}$ is simply a scaled version of the component noise covariance matrix $R_\mathbf{g}$ of (2):

$$R_\mathbf{g}^{(n,m)} = R_\mathbf{g} P^{(n,m)} \qquad (5)$$

The scaling factor $P^{(n,m)}$ is determined by the Gabor filter corresponding to subband (n,m) [6].

Using (4), an expression may be written for the ML estimate of $\hat{\mathbf{f}}(\mathbf{x})$ in each subband (n,m) in terms of the *vector displaced subband difference* analogous to the vector DFD of Konrad and Dubois [3]. However, instead of incorporating neighbouring subpixels to obtain a robust estimate, as in pixel-domain ME, we form a single estimate over the six orientational subbands at $\mathbf{x}$, giving

$$\begin{aligned} \mathbf{f}_0^{(m)}(\mathbf{x}) &= \text{arg min } \left\{SD^{(m)}(\mathbf{x},\mathbf{f})\right\} && (6) \\ \text{where } SD^{(m)}(\mathbf{x},\mathbf{f}) &= \sum_{n=1}^{6} \left(\mathbf{DSD}^{(n,m)}(\mathbf{x},\mathbf{f})\right)^H \frac{R_\mathbf{g}^{-1}}{P^{(n,m)}} \\ &\quad \cdot\mathbf{DSD}^{(n,m)}(\mathbf{x},\mathbf{f}) && (7) \\ \text{and } \mathbf{DSD}^{(n,m)}(\mathbf{x},\mathbf{f}) &= \mathbf{g}_1^{(n,m)}(\mathbf{x}+\mathbf{f}) - \mathbf{g}_2^{(n,m)}(\mathbf{x}) && (8) \end{aligned}$$

is the vector displaced subband difference. This formula relies on the fact that the 6 corresponding Gabor filters have insignificant overlap in the frequency domain, so noise in each subband is approximately uncorrelated [6]. The 6 orientations hence contribute independently to the ML estimate.

The CDWT algorithm forms a quadratic approximation to the matching surface $SD^{(m)}$ by relying on the properties of the underlying Gabor filter to interpolate $g_{1,k}^{(n,m)}(\mathbf{x})$ in between the known integer grid values. The result is

$$SD^{(m)}(\mathbf{x},\mathbf{f}) \approx (\mathbf{f} - \mathbf{f}_0)^T \mathcal{K}(\mathbf{f} - \mathbf{f}_0) + \delta \qquad (9)$$

where $\mathcal{K}$ is the *curvature matrix* of the surface whose parameters may be computed from the subimage coefficients in each colour component.

If $R_\mathbf{g}$ is diagonal, with entries $\{\sigma_1^2, \sigma_2^2, \sigma_3^2\}$, (6) becomes

$$\mathbf{f}_0^{(m)}(\mathbf{x}) = \text{arg min } \sum_{n=1}^{6}\sum_{k=1}^{3} \frac{SD_k^{(n,m)}(\mathbf{x},\mathbf{f})}{P^{(n,m)}\sigma_k^2} \qquad (10)$$

$$\text{where } SD_k^{(n,m)}(\mathbf{x},\mathbf{f}) = \left|g_{1,k}^{(n,m)}(\mathbf{x}+\mathbf{f}) - g_{2,k}^{(n,m)}(\mathbf{x})\right|^2$$

is the *subband squared difference* for subband (n,m) of colour component k. In this case each of the three colour components contributes independently to the ML estimate, inversely weighted by the noise variance in that component. Clearly this is a much simpler case, and the parameters $\mathbf{f}_0, \mathcal{K}$, and δ of $SD^{(m)}$ may be estimated with much less computation than in the general case.

The algorithm proceeds to refine the field of motion estimates (given by $2^{m_{max}}\mathbf{f}_0$) by passing down the surface parameters from coarse to fine (with decreasing m) and incorporating them as prior information to the level below by simple addition of quadratic surfaces. The algorithm halts at level m_{min}, at which the estimate field consists of one estimate per $2^{m_{min}}$ by $2^{m_{min}}$ block of pixels.

2.4 A Decorrelating Transform

In the vector ML formulation of ME, the covariance matrix $R_{\mathbf{g}}$ of the inter-component noise controls the weighting to be assigned to the contributions of the various colour components. If each input 3-component frame is transformed by

$$\mathbf{g}'(\mathbf{x}) = A\mathbf{g}(\mathbf{x}) \qquad (11)$$

where A is an n-by-3 matrix, then the noise covariance matrix in the new colour space ("A-space") becomes

$$R_{\mathbf{g}'} = AR_{\mathbf{g}}A^T \qquad (12)$$

Given some estimate of $R_{\mathbf{g}}$, the singular value decomposition (SVD) of the symmetric, non-negative definite matrix $R_{\mathbf{g}}$ yields orthogonal V and diagonal D such that

$$V^T R_{\mathbf{g}} V = D \qquad (13)$$

Note that the singular values are non-negative; however, one or more may be zero. (This case occurs when at least one component is a linear combination of the others, and it prevents the use of (7) to find the ML estimate, because $R_{\mathbf{g}}$ is non-invertible.) Suppose there are n non-zero singular values (i.e. $R_{\mathbf{g}}$ has rank n). If we extract the invertible n-by-n submatrix D' from D and the corresponding rows V' from V, it can easily be shown that

$$A = (\sqrt{D'})^{-1}(V')^T \qquad (14)$$

is a decorrelating transform for colour space $\mathbf{g}$, i.e. $R_{\mathbf{g}'} = I_n$. The SVD-based transformation thus identifies the case where there are redundant components and projects to a colour space of appropriately reduced dimensionality. Equation 10 may now be invoked, with $\sigma_k^2 = 1$ for all $k = 1, \ldots, n$. Each transformed component contributes equally and independently to the ML estimate.

3 Results and discussion

Synthesised test sequences were obtained by applying motion fields of three distinct kinds—uniform translation, rotation, and divergence—to the 128-by-128 pixel central portions of frame 1 of the "carphone", "foreman", and "suzie" colour sequences respectively. White Gaussian noise of known correlation statistics was added to each frame of the synthetic sequence. The noise covariance matrix in RGB-space, as in [5], was

$$R_{rgb} = \sigma^2 \begin{bmatrix} 1.7393 & 0.1871 & -0.1886 \\ 0.1871 & 0.1318 & -0.0742 \\ -0.1886 & -0.0742 & 0.3654 \end{bmatrix} \qquad (15)$$

The test algorithm was the simplest version of the CDWT algorithm, with $m_{max} = 5$ and $m_{min} = 2$, with no confidence thresholding. A full-density field was obtained by bilinear interpolation from the final motion field. For each of the three test sequences, three sets of results were obtained: those using unweighted RGB-space ME; those from luminance-only estimation; and those from vector ML estimation, obtained by first transforming from RGB to the optimal colour space as described in Section 2.4, using the known R_{rgb}.

Global estimation error was quantified by averaging the angular deviation of the estimated motion fields from the known "true" motion fields in the synthesised sequences, excluding a strip of width 16 pixels around the boundary of the frame. Figure 1 shows the mean error angle of the estimated fields on each test sequence as a function of the noise standard deviation. The results demonstrate that the optimal strategy is the most noise-robust, followed by luminance-only, with the straightforward unweighted combination of RGB fields, ignoring the noise statistics, giving the worst results. Figure 2 shows the lower right quarter of the estimated motion fields for the divergence sequence with $\sigma = 18$, superimposed on images of error angle (darker means greater error.) This figure demonstrates the much improved motion field smoothness gained by using colour optimally. However, the gains are only markedly evident at high noise levels. These results echo those obtained with a pixel-domain gradient-based strategy on the same test sequences [5].

Luminance-only ME requires only slightly more than a third of the amount of computation required for full-colour ME [6]. Our results suggest that this strategy provides near-optimal performance except where noise overwhelms luminance contrast. In such cases, chrominance information may be incorporated (according to the ML formulation) to increase the robustness of the estimates. An adaptive strategy, in which luminance is the primary quantity for estimation, with some criterion to indicate where the chrominance information should be incorporated, would provide the best tradeoff between accuracy and efficiency. Our on-going work is aimed at finding a reliable criterion for incorporation of chrominance information.

Tests on a real colour sequence ("Claire", available in YUV space) were also carried out. An estimate of R_{yuv} was generated from observations of frame differences in stationary areas of the image. Because R_{yuv} is near-diagonal, there was hardly any difference between straightforward unweighted YUV-space ME, and ME based on the optimal use of colour as described in this paper. Furthermore, because the SNR in the y (luminance) component is much

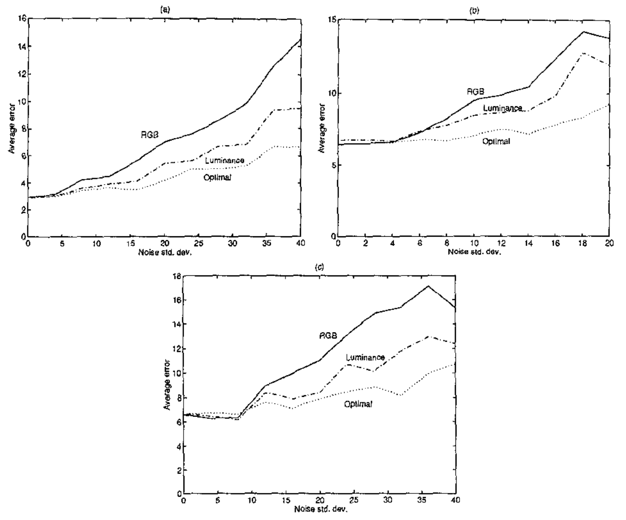

Figure 1: **Mean motion field error vs σ for RGB, luminance-only, and optimal colour estimation. (a) Translation sequence. (b) Divergence sequence. (c) Rotation sequence.**

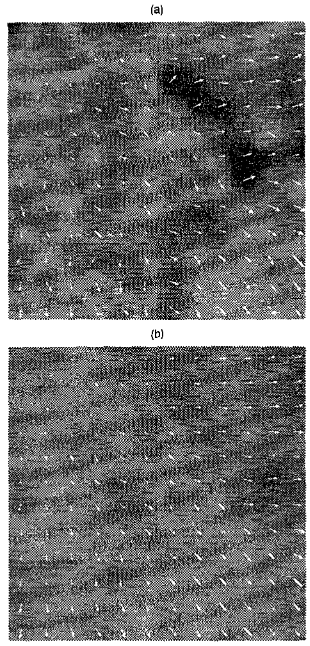

Figure 2: **Lower right quarter of motion fields for divergence sequence with $\sigma = 18$, superimposed on error images (darker means greater error). (a) RGB-estimated field (one estimate per 4 by 4 pixels). (b) Optimally-estimated field (same resolution).**

greater than that of the u and v (chrominance) components, luminance-only estimation gave very similar performance to optimal full-colour estimation. These results, which may be found at our web site[1], further reinforce the arguments for an adaptive scheme.

4 Conclusion

In this paper we have shown how to formulate the ML motion estimator for colour image sequences in the presence of correlated, homogeneous Gaussian noise in the three component fields. The vector ML formulation was applied to the complex-wavelet-domain matching algorithm, with the the inter-component noise covariance matrix playing a pivotal role. We have shown how to use this matrix to derive a transformation into an "optimal" colour space, in which the colour components may be treated as equal and independent contributors to the ML estimate. The effectiveness of the optimal colour space transformation in the wavelet-domain matching algorithm was demonstrated on three synthesised test sequences containing additive noise with deliberately induced covariance. Our tests also showed that luminance-only estimation performs reasonably well by comparison with the more expensive full-colour approach, particularly at low noise levels. This suggests that the best strategy for a general colour sequence would be adaptive.

References

[1] J.F.A. Magarey and N.G. Kingsbury. An improved motion estimation algorithm using complex wavelets. In *Proc. IEEE Int. Conf. on Image Processing*, pages 969–972. IEEE, September 1996.

[2] A. Mitiche, Y.F. Wang, and J.K. Aggarwal. Experiments in computing optical flow with the gradient-based, multiconstraint method. *Pattern Recognition*, 20:173–179, 1987.

[3] J. Konrad and E. Dubois. Use of colour information in Bayesian estimation of 2-D motion. In *Proc. ICASSP*, pages 2205–2208. IEEE, 1990.

[4] B.C. Tom and A.K. Katsaggelos. Resolution enhancement of colour video. In *Proc. EUSIPCO-96*, pages 145–148, September 1996.

[5] J.F.A. Magarey, A.C. Kokaram, and N.G. Kingsbury. Robust motion estimation using chrominance information in colour image sequences. In *Proc. Int. Conf. On Image Analysis and Processing, Florence*, September 1997. *(To appear)*.

[6] J.F.A. Magarey. *Motion estimation using complex wavelets.* PhD thesis, Cambridge University Department of Engineering, 1997.

[1] `http://www-sigproc.eng.cam.ac.uk/~jfam/work/colour/`