

Colour Constrained 4D Flow

Timothy C. Lukins and Robert B. Fisher
School of Informatics
University of Edinburgh
Edinburgh, EH9 3JZ, UK
{s0126209, rbf}@inf.ed.ac.uk

Abstract

The addition of colour information to the computation of range/scene flow is proposed to improve its accuracy and robustness to ambiguities. This is applied in the form of additional optical flow constraints from aligned colour image data. Combining constraints gives improved velocity displacement fields for both synthetic and real datasets over using depth alone, or in using depth plus intensity. This ultimately has benefits for the processing of dense, temporal depth data obtainable from novel video-rate 3D capture systems.

1 Introduction

The computation of a dense instantaneous velocity field that describes the displacements between two or more instances of a surface can be described as the *scene* or *range-flow*. The general distinction between the two terms is that the former is usually derived within a multiple perspective camera framework, while the latter from orthographic scanned data. These both extend the application of optical-flow techniques with the aim to accurately reflect the motion and deformation of surfaces as 4D flow (i.e. across 3D space and time). Recent advances being made in real-time stereo and scanning techniques motivate new research into exploiting this flow-field as a useful tool to further segment, classify and analyse data for a range of applications in the HCI, graphics and medical domains.

Estimates of observed optical-flow can be used to recover depth changes, particularly when using multiple cameras and feature correspondences. As optical-flow is the projection of 3D surface motion onto an image plane, it can be used to recover the *scene-flow* within the context of a physical framework that must take account of the camera's intrinsic parameters and lighting conditions from various angles [12]. Scene-flow can also be computed from raw 2.5D stereo depth-maps as if they were intensity images in order to track the apparent depth change [8]. However, the resulting 2D flow estimates must also be reprojected back into 3D space, potentially leading to error over a large depth-of-field.

Alternatively, an approach can be adopted for raw range data (whether from a stereo or a triangulated sensor) to derive *range-flow*. In theory any optical flow technique can be applied for estimation [5], but the most investigated approach which we directly build on uses differential estimation techniques for measuring the temporal-spatial gradient (i.e. Spies *et al.* [11]). One advantage is that range-flow can be computed directly on the original surface by using localised estimates calculated directly from the sample grid as the basis for creating a regularised, smooth flow field.

Both scene and range-flow are fundamentally similar in that they seek to capture the 4D motion and deformation of a surface. An advantage in using the regularly sampled depth information (e.g. from laser-stripe scanners and dense stereo capture) is that it is often accompanied by aligned intensity or colour information, which may not have been directly utilised in the initial 3D capture. This extra information can further constrain and thereby improve the accuracy, reduce aperture ambiguities, and increase density in the flow estimates - as shown in the case of incorporating intensity [10]. Research in further constraining optical-flow techniques has also investigated the use of texture [2] and colour [1]. Relatively little work has applied these potential sources to range data despite their obvious presence in vision, and in applying the benefits to 4D flow.

The use of colour in particular is proposed on the basis that the information contained within additional colour channels is more representative of the actual surface properties of the objects in question [6]. When considering that intensity depends on more global assumptions about illumination and reflectance, then surface invariant colour characteristics have obvious benefits in resolving certain types of localised motion ambiguity. This is particularly the case in observing non-rigid deformation where attempting to track motion on the surface (even with accurate stereo correspondence) becomes an issue. Recent quantitative assessment of standard optical flow techniques modified to include colour have shown some improvement over simply adding gray-scale intensity [3]. That work has also illustrated the issues in deciding which channels to select and how to adjust the weightings between them.

Thus, there has been research in adding colour to optical-flow [1, 3], and in adding intensity to range-flow [10], yet no work that directly adds colour to 4D-flow estimation. We propose below a generalised approach to incorporating multiple aligned channels based on a simple Least Squares range-flow framework, and show that adding colour can give better accuracy and reduced aperture ambiguity.

2 Calculating Colour Constrained 4D Flow

2.1 Isolating Channels

In our formulation we extend the original work of [10] which considers the varying depth of an orthographically captured surface (X, Y, Z) as a function of time t , locally constraining a *range flow* field in 3D $\vec{f} = [U, V, W]^T$ by:

$$Z_X U + Z_Y V + W + Z_T = 0 \quad (1)$$

where subscripts indicate partial derivatives for the simultaneous change in depth Z with respect to local position in (X, Y, T) . Similarly, since we also possess aligned intensity and colour channels, we can generalise the constraint provided by a *channel C* for observed 3D/2D flow by:

$$C_X U + C_Y V + C_Z W + C_T = 0 \quad (2)$$

where $C_Z = 0$ for colour and intensity channels, and $C_Z = 1$ for range channels in the case of a sensor aligned orthographic projection. Adding these additional constraints can help resolve the *aperture problem* common to optical and range flow, where the normal velocity can be recovered but not the tangential velocity. This can occur for example in the 2D case of a line, and in the 3D case of a fold-edge, for which there is not enough

localised information to resolve motion along the direction of that line or edge. In combining multiple estimates from separate unrelated channels (e.g. range and colour) we anticipate fewer aperture problems due to the combined constraints serving to cancel out such ambiguities. This would lead to an improvement in the overall accuracy.

In integrating colour it is necessary to consider how it can be represented and converted, as defined by CIE standards. In particular, we are interested in isolating the *luminance* component (i.e. brightness, value, intensity) in order to identify the *chromatic* components. Given that our initial input is normal 3-channel *RGB*, one alternative colour space removes the effects of brightness via *normalised RGB* values as:

$$NRGB = \left[\frac{R}{R+G+B}, \frac{G}{R+G+B}, \frac{B}{R+G+B} \right]. \quad (3)$$

We are furthermore interested in the conversion of raw *RGB* to the more useful and accurate device-independent colour space: CIE *LAB*. This is a non-linear transform to a spherical colour-space that directly isolates intensity (*L*) from red-green (*A*) and blue-yellow (*B*) axes. From this it is further possible to convert the space to polar-coordinates, giving the *LCH* representation for chroma (*C*) and hue (*H*) as:

$$H = \arctan\left(\frac{A}{B}\right). \quad (4)$$

Altogether we compare 6 methods of combining additional channel information: depth by itself (*Z*), depth plus raw colour (*ZRGB*), depth and CIE *LAB* (*ZLAB*), depth plus only luminance/intensity (*ZI*), depth plus three-channel brightness invariant colour (*ZNRGB*), and depth plus one-channel hue (*ZH*). For simplicity, we do not consider more device-dependent oriented colour-spaces such as *CMYK*, *YIQ*, and *YUV* (despite being often faster to compute, they are specifically tuned to printing and display requirements).

2.2 Combining Constraints

As with optical-flow algorithms, we wish to locally estimate the motion by solving for $\vec{f} = [U, V, W]^T$. To achieve this further assumptions must be made. One is to assume nearby velocities are equivalent in relative direction and magnitude, and thus use a *Least Squares* minimisation to integrate the velocities into full flow estimates. Here we assume orthographic equally spaced data, but the technique can be adopted for other cases.

Fundamental to our approach is the way in which the individual channels must be combined via a scaling value β . Spies *et al.* [10] calculate a single weighting on the basis of the averages in intensity and depth gradient magnitudes (using a representative training data set, but often only setting the value uniformly to 1.0). They also suggest the possibility of adding multiple channels constraints by simply summing all contributions. We instead desire a more robust and invariant estimation of the contribution by many different channel to the combined estimate.

We do this firstly since we cannot assume that any two channels are in any way related to the same underlying process or sensor capabilities, and so cannot be compared directly in terms of their scale or distributions. Thus, as described in [11] we scale the other channels to have the same mean and variance at the depth data. This enables us to overcome the initial discrepancies in sensor ranges and readings.

Secondly, we wish to give precedence to channels that are more reliable over those that are worse affected by noise and by aperture ambiguity. A Lucas and Kanade [7, 5, 11] based Least Squares approach can incorporate weighting when constructing a system of linear equations to solve \vec{f} for a local neighbourhood of size N pixels, over M channels:

$$\underbrace{\begin{bmatrix} \sum_{e=1}^M \beta_e C_{X_1} & \sum_{e=1}^M \beta_e C_{Y_1} & 1 \\ \vdots & \vdots & \vdots \\ \sum_{e=1}^M \beta_e C_{X_N} & \sum_{e=1}^M \beta_e C_{Y_N} & 1 \end{bmatrix}}_A \vec{f} = \underbrace{\begin{bmatrix} -\sum_{e=1}^M \beta_e C_{T_1} \\ \vdots \\ -\sum_{e=1}^M \beta_e C_{T_N} \end{bmatrix}}_B. \quad (5)$$

In this way we are simply seeking to weight and combine the information provided by all the channels within the same framework, as dictated by their β values. Each individual channel's coefficient matrix for the local planar 2D neighbourhood:

$$\begin{bmatrix} (\sum C_X^2) & (\sum C_X C_Y) \\ (\sum C_X C_Y) & (\sum C_Y^2) \end{bmatrix} \quad (6)$$

can be constructed and assessed for its reliability ρ as the *reciprocal of the condition number* derived from the ratio of its maximum to minimum eigenvalues:

$$\rho = \frac{1}{\lambda_{max}/\lambda_{min}}. \quad (7)$$

This represents the numerical stability of that channel's contribution to the Least Squares calculation. The beta value then expresses this contribution as a weighting for each channel in the neighbourhood over the summation of all other channel reliabilities:

$$\beta_e = \frac{\rho_e}{\sum \rho}. \quad (8)$$

If the summation $\sum \rho$ is not greater than a threshold level θ , we can reject outright the estimation for this neighbourhood as ill-conditioned.

2.3 Deriving Flow

To numerically generate a flow estimate for each uniformly spaced pixel of the aligned input channels we construct the respective matrices A, B as defined above. We use robust balanced Simoncelli filters [9] for derivative estimation which have been shown to provide the best results by employing two stage low pass (noise reduction) and high pass (differentiate) 5-tap filters. These are applied as convolutions across the appropriate dimensions respectively for each channels C_X, C_Y and C_T . For example, to calculate C_X we first convolve across the T dimension using the smoothing kernel $[0.036, 0.249, 0.431, 0.249, 0.036]$, followed by similar smoothing across the Y dimension, before finally convolving the differentiation kernel $[-0.108, -0.283, 0.0, 0.283, 0.108]$ across the X dimension.

For Least Squares, the flow estimate can then be computed from these combined derivatives by the pseudo inverse of equation 5:

$$\vec{f} = (A^\top W^2 A)^{-1} A^\top W B \quad (9)$$

where the W component expresses a weighting matrix over the neighbourhood N drawn from a zero mean 2D Gaussian. We also use a threshold of $\theta = 0.5$ to reject estimation.

3 Experiments

3.1 Synthetic Data

To quantitatively test the benefits of additional constraint channels we assess their accuracy in predicting the known rigid displacement of a surface. The two orthogonal synthetic data-sets we rely on are a sloped plane (“slope”) and a sinusoidal plaid pattern (“splaid”), each of a sampling size 100×100 as shown in figure 1.

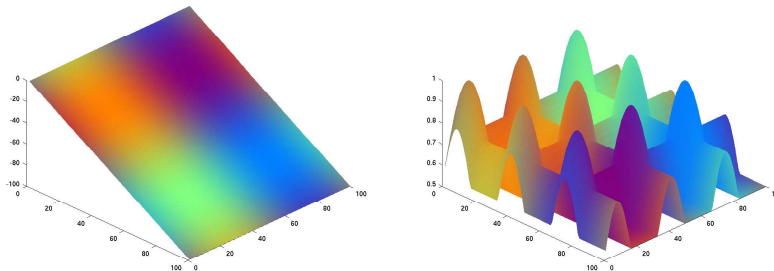


Figure 1: Synthetic slope and splaid data-sets with depth and RGB colour.

The splaid surface is generated over the functional space of $2\pi \times 2\pi$ for a frequency of 3 (that is, repeated every 32 pixels). Of critical importance to these experiments is that the colour and depth values are generated over the same functional space, such that the observed displacement in depth must match exactly the motion of colour values on the surface. We also seek to create a varied pattern that contains different aperture ambiguities when represented in various colour channels, generated again by a interwoven combination of sine waves - as seen in figure 2. Noise can also then be added by a random amount drawn from a Gaussian distribution of standard deviation σ separately to each input channel (representing the deficiencies in sensor capabilities).



Figure 2: Colour data in RGB, Intensity, Normalised RGB, LAB, and Hue.

To assess accuracy we use the standard [4] measurements of the relative *magnitude error* (E_m) as a percentage, and *angular error* (E_a) in degrees between estimate f_{est} and known correct displacement f_{cor} as:

$$E_m = \frac{\| \|f_{est}\| - \|f_{cor}\| \|}{\|f_{cor}\|} \times 100, \quad E_a = \arccos\left(\frac{f_{cor}}{\|f_{cor}\|} \cdot \frac{f_{est}}{\|f_{est}\|}\right).$$

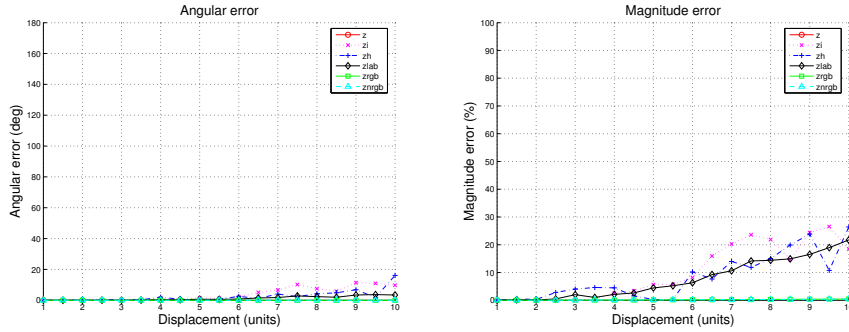


Figure 3: E_a and E_m for translation ($1 \leq T_X \leq 10$) of the slope dataset.

Using these, we take the mean error over all the estimated flow vectors, and observe the effects of 1D translation ($1 \leq T_X \leq 10$) of the slope dataset as shown in figure 3 where $1 \text{ pixel} \approx 1 \text{ unit}$. These show a complete failure to derive an estimate based on Z alone due to the aperture ambiguity of the slope creating an ill-conditioned solution that is rejected. However, combining additional information from the other channels allows this to be resolved. These estimates progressively worsen as the plane translates further across X , yet the RGB and NRGB solutions consistently provide the overall best results. The addition of LAB also provides a relatively robust solution, while the addition of H erratically affects the estimation due to its polar representation of colour which can result in sudden transitions in spatial gradients.

The effects of a more complex uniform translation in 3D ($1 \leq T_X, T_Y, T_Z \leq 10$) for the surface of the sploid dataset are shown in figure 4. Here the Z channel has enough local information from the surface to make a good estimate and perform well for estimation of smaller translations (especially in angle). Interestingly, the combined use of RGB and NRGB has little improvement as they have an overall lower reliability than the Z estimate. However, the addition of LAB colour provides the most robust correction for larger translations due to the A and B channels having a relatively higher reliability.

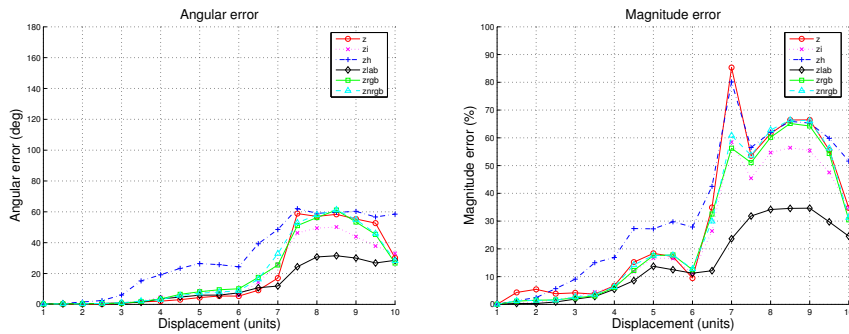


Figure 4: E_a and E_m for translation ($1 \leq T_X, T_Y, T_Z \leq 10$) of the sploid dataset.

The addition of varying levels of noise to the slope dataset produce consistently inferior results for NRGB and RGB as shown in figure 5 for the magnitude error of the slope dataset (again translating only in X). The combined Intensity and LAB perform better in this experiment due to the noise smoothing that occurs when combining separate channel RGB input data. If the intensity channel were from a true intensity sensor, the effective noise level would be higher and thus results would be worse.

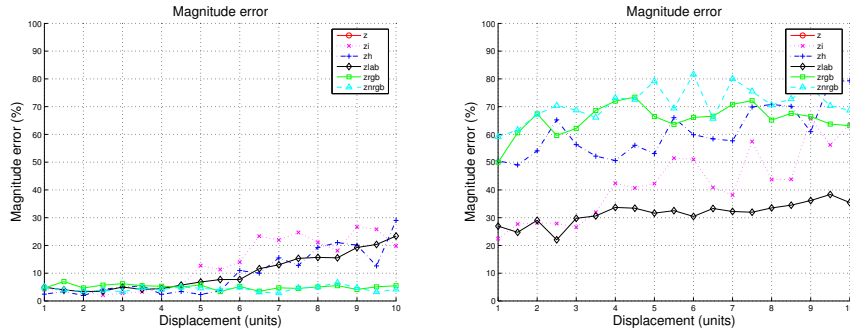


Figure 5: E_m for noise $\sigma = 0.001$ (left) and $\sigma = 0.01$ (right).

3.2 Real Data

As a more qualitative assessment of the accuracy of the colour enhancement, we present work on using 4D scene flow to look at the motion for an expression occurring on the human face. A stereo capture rig constructed of two cameras was calibrated and used in burst mode to capture a sequence of a subject making a “surprised” expression at 2.5 frames per second (each pixel $\approx 2mm^2$). Dense stereo data was recovered from each of the two simultaneous images via stereo photogrammetry, and transformed back to the coordinate frame of the original left image for aligned colour data - shown in figure 6. For computation, and to avoid aliasing, we reduce the sampling resolution to 80×120 .

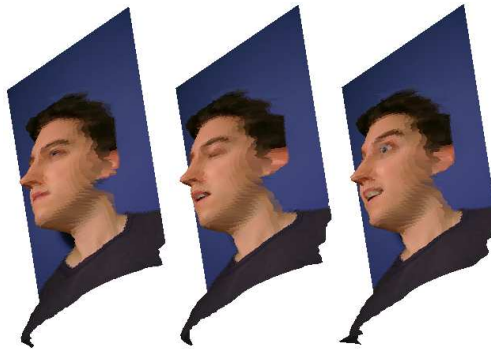


Figure 6: Real data “surprise” sequence frames.

Using our Least Squares approach we derive a range flow estimate between successive frames of data, incorporating additional channels as shown in figure 7.

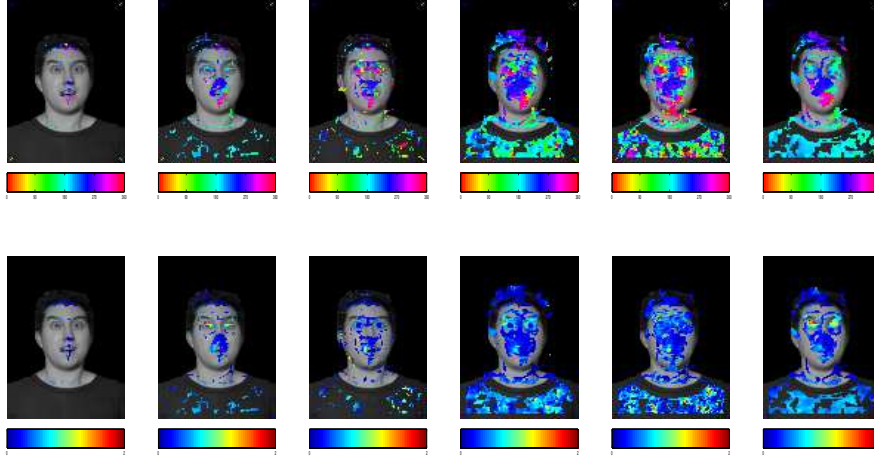


Figure 7: Angle (top) with $0^\circ = \text{down} = \text{red}$, and magnitude (bottom) estimates over subsequent frames for Z, ZI, ZH, ZLAB, ZNRGB, and ZRGB (left to right).

As can be seen, the additional channels immediately lead to an improved density estimation compared to the Z channel alone. However, this can vary considerably due to the large amounts of noise, and can in consequence predict too much planar displacement. Overall, the ZLAB solution appears to capture best the motion of the eyes and jaw. This is confirmed by comparing the Sum Squared Differences (SSD) between the actual surface displacement in the two original frames of depth data, and the predicted displacement of the range-flow (i.e. warping the previous frame) shown in table 1.

Channels:	Z	ZI	ZH	ZLAB	ZNRGB	ZRGB
SSD	0.110679	0.111088	0.111171	0.112795	0.109312	0.117954

Table 1: Sum Square Difference between warped and actual surfaces.

4 Conclusion

The addition of colour (or intensity) constraints within the range-flow calculation can be shown to improve the estimation and resulting flow, as opposed to using depth alone, by reducing aperture errors. In particular, the combined use of raw RGB channels can generally reduce error further than simply combining with Intensity. The use of more compact colour representations do in general provide help in resolving ambiguities, and can prove robust in the presence of greater levels of noise. More channels also lead to greater density of estimation.

We plan to use the increased accuracy offered by these extensions for the possibility of analysing data obtainable from the next generation of video-rate 3D capture systems.

Acknowledgements

This work is supported by an Imaging Faraday CASE award in conjunction with Dimensional Imaging Ltd. (www.di3d.com).

References

- [1] R.J. Andrews and B.C. Lovell. Color optical flow. In B.C. Lovell and A.J. Maeder, editors, *Proceedings Workshop on Digital Image Computing*, pages 135–139, 2003.
- [2] M.A. Arredondo, K. Lebart, and D. Lane. Optical flow using textures. *Pattern Recognition Letters*, 25(4):449–457, 2004.
- [3] J. Barron and R. Klette. Quantitative color optical flow. In *ICPR*, volume 4, pages 251–255, 2002.
- [4] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.
- [5] S.S. Beauchemin and J.L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466, 1995.
- [6] P. Golland and A.M. Bruckstein. Motion from colour. *Computer Vision and Image Understanding*, 68(3):346–362, 1997.
- [7] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *7th Int. Conf. on Artificial Intelligence*, pages 674–679, 1981.
- [8] J.-C. Nebel and A. Sibiryakov. Range flow from stereo-temporal matching: Application to skinning. In *IASTED Int. Conf. on Visualization, Imaging, and Image Processing*, 2002.
- [9] E.P. Simoncelli. Design of multi-dimensional derivative filters. In *IEEE Int. Conf. Image Processing*, volume 1, pages 790–793, 1994.
- [10] H. Spies, B. Jähne, and J. Barron. Dense range flow from depth and intensity data. In *ICPR*, pages 131–134, 2000.
- [11] H. Spies, B. Jähne, and J. L. Barron. Range flow estimation. *Computer Vision Image Understanding*, 85(3):209–231, 2002.
- [12] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), 2005.