Robust Differential Invariant Extraction for Scene Reconstruction

Shih Ching Fu and Peter Kovesi
School of Computer Science & Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley, W.A. 6009, Australia.
email: {scfu, pk}@csse.uwa.edu.au

ABSTRACT
This paper presents the results of using a correlative filtering method for extracting the divergence and deformation of a motion field for later use in scene reconstruction. The motion field created by a camera in motion encodes useful structural information about the scene it is capturing, but these motion fields are rarely recovered perfectly and are often corrupted by noise. It is hypothesised that extracting the divergence and deformation of an optical flow field using a filtering method will be robust to noise present in that field. A comparison is made between the noise robustness of using the first-order differentials determined analytically to those derived via filtering. The results show that a filtering method is more noise robust but much improvements need to be made to improve the localisation of the estimates.

Keywords: Structure from motion, vector fields, image divergence, image deformation.
CR Classification: I.2.10 [Vision and Scene Understanding]: Motion

1. INTRODUCTION
Much of Computer Vision is concerned with deriving three-dimensional (3D) scene information from two-dimensional (2D) images. The aim is to replicate the human visual system which can discern quantities such as depth and motion, allowing interaction with the environment without needing to come in direct contact with it [6]. Of particular interest is studying the role that visual motion plays in gathering information about our surroundings. In the absence of stereo, a moving monocular observer can determine scene structure due to the disparities between successive viewpoints. Using such cues we can formulate solutions to the structure-from-motion problem.

The focus of this paper is on the analysis of image motion or optical flow, and in particular the extraction of the first-order differential invariants of image velocity using correlative filtering. While much literature has dealt with the robust and accurate extraction of optical flow and using this data to directly solve for scene shape, much less literature has examined the use of the first-order derivatives of optical flow. Of the work that has examined the use of the first order quantities of divergence, curl, and deformation, the noise sensitivity of these quantities is an obstacle to accurately calculating surface estimates. Rather than try to further improve the output of optical flow algorithms, we propose a method of extracting these first order derivatives which is much more tolerant of imperfect data.

2. DIFFERENTIAL INVARIANTS OF IMAGE MOTION
Koenderink and Van Doorn [7, 8] have shown how the distortion of an object’s 2D image can be decomposed into components of divergence, curl, and deformation. A geometric summary of these is provided in Fig 1. These quantities
are known as the first-order differential invariants of image motion. They are particularly useful for scene reconstruction because they can be expressed in a coordinate system independent manner. Cipolla [5] provides an excellent summary of the properties which Koenderink proposes and in particular how these quantities relate to the structure of a scene.

2.1 Dilatation
Dilatation is interpreted geometrically as isotropic expansion about a focus of divergence, or, conversely, isotropic contraction about a focus of convergence. It is typically due to the relative motion of objects toward or away from the camera. The relationship between the divergence at a point in the vector field and the orientation of the tangent plane at that point is discussed in the next section. In addition to encoding some surface orientation information the divergence also provides some temporal depth information in the form of time-to-contact.

Time-to-contact, \( t_c \), is a useful quantity for obstacle avoidance problems and can be easily calculated for scenes containing pure divergence. Equation 1 is used to calculate depth maps measured in temporal units (frames to impact) rather than spatial units.

\[
t_c = \frac{2}{\text{div} \, \vec{v}}
\]

2.2 Vorticity
Vorticity corresponds to 2D rotation of an object’s image as shown by the top left illustration in Figure 1. It is expressed as a curl component in the motion field and is orthogonal to the divergence component. The curl component of motion fields is a very problematic quantity since viewer rotations often cancel out the effects of viewer translation. For example, an observer who pitches up to track a rising object will nullify the expected upwards optical flow. Maybank [9] poses some solutions to this problem.

2.3 Shearing
Deformation or shearing, while affecting edge orientations, does not change the apparent area of a closed contour. Deformation fields are characterised by an axis of maximum extension and another perpendicular axis that describes the direction of maximum contraction. Therefore two components, a symmetric component and an anti-symmetric component, are needed to sufficiently define the shear distortion of an image shape. The two components of deformation are illustrated in the bottom two diagrams of Figure 1. As discussed later, the deformation of a motion field provides the most information about scene structure and is not contaminated by other quantities such as time-to-contact and is unaffected by camera rotation. Note that only the magnitude of the deformation is a differential invariant.

2.4 Scene structure from divergence and deformation
The above-mentioned properties are directly related to 3D scene structure and ego-motion, and can be deduced from their effect on scene geometry. In particular, divergence relates to dilatation; curl relates to vorticity; and deformation relates to symmetric and anti-symmetric shearing. In this section we outline how information about surface orientations and time-to-contact can be extracted from these quantities. More specifically the slant, \( \sigma \), and tilt, \( \tau \), of the surfaces in an image are found. We have adopted the notation used by Cipolla [4].

\[
\| \text{curl} \, \vec{v} \| = -\Omega \cdot \vec{Q} + \| \vec{F} \times \vec{A} \| \tag{2}
\]
\[
\text{div} \, \vec{v} = \frac{2 \vec{U} \cdot \vec{Q}}{\lambda} + \vec{F} \cdot \vec{A} \tag{3}
\]
\[
\text{def} \, \vec{v} = \| \vec{F} \| \| \vec{A} \| \tag{4}
\]
\( \vec{v} \) is the image velocity field, \( \vec{U} \) is the translational velocity of the camera, \( \Omega \) is the rotational velocity of the camera, \( \vec{Q} \) is a unit vector defining the viewing direction, and \( \lambda \) is the distance to the object along \( \vec{Q} \). Cipolla uses \( \vec{A} \) to signify the component of the viewer’s translational velocity parallel to the image plane scaled by depth and \( \vec{F} \) to denote the depth gradient scaled by depth.

\[
\vec{A} = \vec{U} - \left( \frac{\vec{U} \cdot \vec{Q} \vec{Q}}{\lambda} \right) \tag{5}
\]
\[
\vec{F} = \frac{\nabla \lambda}{\lambda} \tag{6}
\]

The axis of maximum extension is related to the deformation of an object’s apparent shape and is useful for defining an origin for measuring tilt and is denoted \( \mu \).

\[
\mu = \frac{\angle A + \angle F}{2} \quad \text{def} \, + \vec{v} \quad \text{def} \, - \vec{v} \tag{7}
\]

Where \( \text{def} \, + \vec{v} \) and \( \text{def} \, - \vec{v} \) denote the symmetric and anti-symmetric components of deformation respectively.

Finally, the relationship between \( \vec{F} \) and the slant and tilt is given by Equations 9 and 10.

\[
\sigma = \arctan \| \vec{F} \| \tag{9}
\]
\[
\tau = \angle \vec{F} \tag{10}
\]

Notice that all the scene structure information is encoded into \( \vec{F} \), the depth gradient, and that the magnitude of the deformation is sufficient to calculate the surface slant so long as viewer translation is known (as shown in Equation 4).

3. EXTRACTION OF DIFFERENTIAL INVARIANTS
Cipolla and Blake [3, 5] have done work on how to derive surface orientation and time-to-contact from divergence and deformation information. They use B-spline snakes to track the change in the apparent area of scene objects to approximate the divergence and deformation of the motion field. However, this method proves problematic in the absence of trackable features. Work has also been done by Nelson and Aloimonos [10] on the use of divergence for obstacle avoidance; deriving it mathematically but needing to use many images over time to produce good results. The time-to-crash detector implemented by Ancona [1] utilizes optical flow directly rather than post-processing for divergence. Most of these approaches have only treated the divergence of the optical flow field and avoided using the deformation due to its
high noise sensitivity. Unfortunately, this means discarding much of the scene structure information.

The remainder of this paper discusses the possibility of using a correlative filtering method to extract divergence, curl, and deformation from dense motion fields.

3.1 Filter Design
Analogous to how image features can be detected by using tuned filters, it is proposed that a vector field can be similarly processed for dilatation, vorticity, and shearing. Below we present a preliminary design for filters or detector masks for recovering divergence, curl, and deformation from optical flow fields. Examples of such filters are shown in Figures 2 to 5.

Each filter is a simple vector field depicting divergence, curl, or deformation as appropriate. The magnitude of the individual vectors is proportional to its distance from the centre of the filter. Previous work has included a modulation of the vector magnitudes by a 2D Gaussian envelope to reduce the possible effects of a sharp discontinuity at the mask boundaries. However, this was shown to distort the filter responses and has been replaced by circular step-edge envelope.

The direction of the vectors in the mask depends on the mask type. Divergence masks have vectors that point radially outward as seen in Figure 2, curl masks have tangential vectors as depicted in Figure 3, and the vectors in the deformation masks lie between the axes of maximal expansion and contraction. For deformation, the two cases necessary to capture distortion in all directions are distinguished by the tilt of their axis of maximum extension, $\mu$, as illustrated in Figures 4 and 5. The size of the filter mask will affect the size of the features that are detected. Currently the size of filter masks is selected manually with further examination necessary to determine a method for automatically generating a bank of appropriately sized filters.

4. EXPERIMENTAL RESULTS
The rest of this paper discusses the experimental tests we conducted to compare the noise robustness of analytically deriving the first-order differentials with a filtered method for deriving.

Note that in this analysis it is assumed that there is no cam-
era rotation. The details of the curl of a motion field was outlined above for completeness but will be omitted from the rest of the analysis. Also, although the equations presented by Cipolla assume a spherical camera model, we have chosen to approximate this using a planar image surface. It is anticipated that distortion will occur on points far from the image centre.

4.1 Ground truth Data

Figures 7, 6, and 8, depict the test scene to be reconstructed as viewed by the camera. The camera has a field of view of $90^\circ$ and a focal length, $f$, of 1 unit. The images are of dimension 300 x 400 pixels using a perspective camera model.

The test scene comprises a textured cube of side length 100 units centred at $(0,0,100)$ with one face removed to allow the camera to capture its interior. The camera is initially positioned at the origin, $(0,0,0)$, and then translated along the camera’s principal axis to $(0,0,10)$. Such a scene will provide an uncontaminated estimate for the time-to-contact using the divergence of the far face of the cube as well as sufficient deformation of the side faces to provide surface orientation information. The texture of the cube as shown in Figure 6 is for illustrative purposes only and not used in optical flow extraction. The motion fields in these experiments have been computed using point correspondences between successive frames.

4.2 Noise free motion fields

Given ground truth data that is noise free, the divergence and deformation of a motion field can be accurately calculated analytically. The magnitudes of the ‘clean’ analytic divergence and deformation of the motion field depicted in Figure 6 is shown in Figures 9 and 11 respectively. Note that although there are two orthogonal components of deformation corresponding to perpendicular shear directions, only the combined magnitudes of these components is required for surface orientation recovery.

In Figure 9 we can see that the back face of the cube can be easily segmented due to the low divergence of its motion field relative to the divergence of the side faces. This contrast is expected since from Equation 3 it is clear that the larger the angle between the surface normal at a point and the translational velocity at the same point ($\angle \mathbf{A} - \mathbf{r}$), the greater the divergence. For the side faces of the cube whose velocity towards the camera is close to perpendicular to the camera translation the divergence is high. For the back face where the translation and surface normals are close to parallel, the divergence is much less.

The divergence calculated over the square back plane is 0.1429. Using Equation 1 we find the expected time-to-contact to be $2/0.1429 = 14.00$ frames. This value matches the expected time given of contact for approaching a surface that is 140 units away at 10 units/frame. A cross-section showing the progression of divergence values horizontally over Figure 9 is shown in Figure 10.

Figure 11 shows the deformation computed analytically over the original field (Figure 6). Although there are two com-

Figure 6: The true motion field superposed over the test scene. The camera has translated by 10 units along the positive Z axis.

Figure 7: The true depth of the scene measured in units from the camera centre.

Figure 8: The true surface orientations of the test scene. Note that four surfaces are parallel to the camera axis and one is perpendicular to it.
components to deformation, it is the overall magnitude that is invariant and used for scene reconstruction (see Equation 4). This is calculated using a simple square root of the sum of squares. Notice that the back plane of the cube exhibits no deformation since its apparent shape does not change, only its area. Figure 12 shows this very clearly by the zero value of deformation across the centre of the image.

Once values for the deformation are extracted, Equations 4, 7, and 8 can be used to compute $|F|$. We approximate $|A|$ as the optical flow scaled by the pixel size. Once a value for the depth gradient is obtained we can use Equations 9 and 10 to recover the slant and tilt of the surface. The results of this reconstruction are shown in Figure 13.

As an initial comparison, the reconstruction is performed again but using the divergence and deformation responses from filter masks similar to those in Figures 2, 4, and 5. The mask size used was $19 	imes 19$ pixels where the deformation detectors had $\mu = 0^\circ$ and $\mu = 45^\circ$. A numerical comparison of the slants and tilts from these two reconstructions shows negligible difference. This goes to confirm the accuracy of a filtering method for extracting differential invariants and provides a good base of comparison when noisy optical flow is considered.

4.3 Noise corrupted motion fields
In practice it is very unlikely that optical flow can be perfectly extracted from an image sequence. Given that divergence and deformation are first-order derivatives they are highly sensitive to noise and is why they are unattractive quantities to use for scene reconstruction. It is well known that filtering procedures are more robust to noise at the expense of localisation and this section will examine whether such robustness extends to divergence and deformation extraction.

In this experiment, the motion field shown in Figure 6 has normally distributed noise of mean zero and standard deviation of 4 added to it. The resultant vector field, subsampled
by a factor of 10 is shown in Figure 15. By inspection, there appears to be only subtle differences between the noisy field and the original ‘clean’ field. However, the subsequent values for divergence and deformation are in great contrast to the originals.

The magnitude of deformation calculated analytically is illustrated in Figure 16 and as expected, it too noisy for discerning any scene structure. Correlation with a deformation detectors produced a much more useful output as seen in Figure 17. To the astute observer it is still possible to identify the different surfaces of the cube as well as the discontinuities where each face intersect.

Given the garbled deformation values from the analytic calculations, the extraction of surface normals from these values is equally as inaccurate. Figure 18 shows the mess resulting from using Cipolla’s equations given noisy first-order data. The recovery of surface normals from the filtered deformation responses is much better, particularly for the side faces of the cube (Figure 19). The centre of the image is still very poor but expected since from Figure 12 we know that there is zero deformation there; any additive noise will markedly affect the calculations. Further tests using larger filter masks improved the robustness but reduced the localisation of the cube’s internal edges. Note that the edge artifacts introduced by filtering have been truncated from the images to improve visualisation.

4.4 Smoothed Motion Fields

Rudimentary smoothing by a Gaussian mask is often used to reduce the effects of noise in data. The next experiment examines the effect of smoothing a noise corrupted motion field before extracting the divergence and deformation. The noisy data (as depicted in Figure 15) is first smoothed by a Gaussian filter before analytic differentiation and correlative filtering are applied for deriving the differential invariants.

A comparison of the two methods of deformation extraction shows that the a filtering method, even after Gaussian smoothing, provides a cleaner result. This is expected since
Figure 16: The analytically derived deformation of the field shown in Figure 15. Notice that it is completely useless for scene recovery.

Figure 17: The filter response of the deformation masks with radius 30 pixels to the noisy vector field. It proves to be much cleaner than the analytic result.

Figure 18: The needle diagram using the analytically derived deformation in Figure 16. No scene structure is discernible.

Figure 19: The needle diagram using the correlated values for deformation. It is an improvement over Figure 18 but the reduced localisation of the filtering method means the intersections of the cube’s surfaces are poorly defined.

any kind of filter mask will tend to even out any local anomalies. Such anomalies are still visible in the analytic response as shown in Figure 20. Although the filtered response in Figure 21 does not have these noise artifacts, there is some distortion near intersection of the cube’s faces. These distortions are a problem because they will skew the computed surface normals away from the true values. This is perhaps a consequence of decreased localisation with increased in filter mask size.

The surface normal extraction from the deformation information confirms that a filtering method will distort the needle diagram around surface discontinuities. However, only one filter size has been considered here (radius = 30 pixels) and perhaps the superposition of the responses of differently sized filter masks will correct this distortion. Further more, the use of morphological techniques may be able to remove the edging artifacts caused by mask correlation.

5. CONCLUSION
The first step in solving a structure-from-motion problem is to compute the optical flow between subsequent images. This process is very difficult and often the results are inaccurate or incomplete. Consequently, the differentials of such data are even more error prone and not even considered for analysis unless better or ‘cleaner’ data can be obtained. In this paper we have proposed a method of more robustly estimating these differentials of divergence, curl, and deformation. This builds on the established knowledge that these quantities are useful for scene reconstruction.

The results show that the use of mask correlation for extracting divergence and deformation is quite robust. In our experiments we compared the response of this correlative method with analytically derived differentials over a noisy motion field. This experiment was repeated after Gaussian smoothing was applied and produced the same conclusion that the filtering method is more robust. These conclusions were made by examining the resultant scene reconstructions
Figure 20: After being smoothed by a Gaussian mask, the analytical calculation of the deformation is much improved from Figure 16 but still exhibits localised anomalies that will hinder accurate scene reconstruction.

Figure 21: The deformation of the same smoothed motion field of Figure 20 using deformation detectors of radius 30. The response is much smoother but at the expense of sharp localisation.

Figure 22: The reconstruction of the cube scene using the deformation information of Figure 20. The surface normals for the cube’s back face are mostly incorrect.

Figure 23: The reconstruction of the cube scene using the deformation information of Figure 21. The surface normals are much improved over the analytically derived deformation, particularly where the signal to noise ratio is greater (along the back face).
and comparing these with the known truth.

However, the results show the drawback that improved noise robustness has come at the expense of accurate localisation of surface discontinuities.

6. FUTURE WORK

The motion field data used in our experiments was dense and uniform in nature. Further investigation needs to be made into the use of incomplete data since this will bias the filter responses. Methods for interpolating the optical flow or even the differentials are needed to fill in possible gaps in data.

The use of morphological techniques for eradicating the edging and smoothing artifacts of filtering needs to be investigated. Post processing the responses of the filters will be necessary to improve the localization of depth discontinuities and use of a bank of differently sized filters will perhaps fix this problem.

Concern also arises over the use of Cipolla’s spherically derived equations but with a planar perspective camera model. The distortion of approximating a spherical projection with a planar one may adversely affect the extracted deformation values and hence the accuracy of the final reconstructed scene. All these issues need to be addressed.

7. ACKNOWLEDGEMENTS

This work is supported by an Australian Postgraduate Award and an ad hoc scholarship from the School of Computer Science & Software Engineering at The University of Western Australia. Thanks goes to my supervisor Peter Kovesi for his assistance and guidance.

8. REFERENCES


