Scene Flow Estimation using Intelligent Cost Functions

Simon Hadfield
S.Hadfield@surrey.ac.uk
Richard Bowden
R.Bowden@surrey.ac.uk

Centre for Vision Speech and Signal Processing
University of Surrey
Surrey, UK

Abstract

Motion estimation algorithms are typically based upon the assumption of brightness constancy or related assumptions such as gradient constancy. This manuscript evaluates several common cost functions from the motion estimation literature, which embody these assumptions. We demonstrate that such assumptions break for real world data, and the functions are therefore unsuitable. We propose a simple solution, which significantly increases the discriminatory ability of the metric, by learning a nonlinear relationship using techniques from machine learning. Furthermore, we demonstrate how context and a nonlinear combination of metrics, can provide additional gains, and demonstrating a 44% improvement in the performance of a state of the art scene flow estimation technique. In addition, smaller gains of 20% are demonstrated in optical flow estimation tasks.

1 Introduction

Scene flow is the 3D counterpart to optical flow, describing the 3D motion field of a scene, independent of the cameras which view it. This paper presents a simple solution to one of the fundamental limitations in scene flow estimation, that of non-conformity to the underlying assumption of brightness constancy. Motion estimation is a fundamental tool in computer vision. Its 2D variant (optical flow) has long been studied, with multiple variants included in most vision libraries. It forms the basis or pre-processing step for many other algorithms. Estimation of scene flow is likely to become equally important in the future, given the rise of commercial depth sensors and 3D broadcast footage.

However, dense motion estimation is difficult. It is well known that there are many situations in which the brightness constancy and related assumptions are invalid [33]. These include non-Lambertian surfaces, occlusions, non uniform lighting and moving light sources. Optical flow estimation attempts to mitigate these issues using separate subsystems (such as occlusion estimation [30] and non local filtering [38]) to detect and ignore certain classes of artifacts. However, optical flow has the fundamental advantage that when estimating the motion between frames, both images come from the same sensor. In scene flow estimation these issues are exacerbated by the use of multiple sensors, with different response characteristics.

In this paper, an in depth analysis of many commonly employed motion estimation metrics is performed. A new "Intelligent Cost Function" (ICF) is then proposed, employing machine learning techniques, and providing improved robustness. These Intelligent Cost Functions (ICFs) are shown to provide performance gains of 44% in a state of the art scene flow framework. This is possible, as the metric learns to penalize brightness inconsistencies

which are indicative of motion estimation errors, while being less sensitive to scene artifacts such as specularities.

2 Related Work

Motion estimation techniques can generally be separated into 3 categories. The first is local approaches [18, 20, 42], inspired by the work of Lucas and Kanade [22], which estimate motion independently at each point, making use of local contextual information. Generally in such approaches, a patch from the source image(s) is matched with a patch from the target image(s), often with an intermediate warping step. This warping accounts for patch deformations, due to motions which are not fronto-parallel. Such techniques are particularly suitable for sparse motion estimation tasks, such as tracking a subset of salient points. However, local approaches provide poor performance in areas of the scene with little texture, as there is insufficient contextual information to obtain a unique match. The second, and arguably more common, category of motion estimation techniques are variational approaches [6, 17, 27, 36], inspired by the work of Horn and Schunck [16]. These techniques perform a global optimization for the motion across the entire scene, with regularization based on the total variation within the estimate. As a result, such techniques are well suited to dense motion estimation tasks, and are able to "fill in" the motion in untextured regions, by smoothly interpolating from the boundaries, so as to minimize the total variation. The third paradigm in motion estimation is sampling based approaches, which are most frequently encountered in the scene flow estimation literature [5, 14, 29, 37] due to the computational complexity of variational approaches in the higher dimensional 3D estimation task. Unlike local and variational approaches, such schemes generally do not involve an optimization stage in order to move along the motion field energy surface. Instead the energy surface is densely sampled and a subset of consistent samples extracted.

One element is common to all these approaches, the encoding of constancy assumptions within a validity metric. Depending on the motion estimation scheme, these metrics serve to either guide energy minimization (locally or globally) or to identify samples which are valid. There has been a rich history in the field of optical flow estimation, exploring various assumptions and related metrics. However, very little such work has been done for scene flow, involving multiple sensors. The developments of optical flow metrics began with a number of simple re-formulations of the initial quadratic cost of [16], including the l_1 norm and robust variants such as the Charbonnier function [8, 9, 39] and the Gaussian Scale Mixture model of Sun *et al.* [33]. Brox *et al.* then proposed supplementing the brightness constancy assumption with gradient constancy [8], while Xu *et al.* utilize one of these two assumptions on a per pixel basis [41]. Kim *et al.* [19] learn a cost function based on the weighted combination of other standard cost function, while Sun *et al.* extended this idea to response constancy, for a small number of 3×3 linear filters [33]. One major limitation of these previous works (highlighted in section 3) is that analysis is limited to visual error statistics of true motion fields (i.e. modelling of scene artifacts). No analysis is performed on the behaviour of these statistics, when the motion field is incorrectly estimated (which in practice is when the matching metrics are most needed). Intuitively, this accurate modelling of ground truth visual consistency ensures that correct motion fields are always recognised as such (reducing "False negatives"), but it tells us nothing about the metrics ability to reject erroneous motion fields ("False positives"). This is likely due to the decreased severity of the issue in traditional optical flow scenarios, where sensor responses are at least consistent.

There has been much work focused on mitigating, rather than addressing, the shortcom-

ings of these existing data matching metrics. Aodha *et al.* [24] developed an approach to recognise the most vital areas to employ these mitigation techniques, The most common of which are multi-scale coarse-to-fine estimation and iterative warping schemes [1, 8, 13, 17]. These are designed to avoid the local minima that are ubiquitous due to the extreme non-convexity of the energy surfaces. In addition, "structure-texture decomposition" [2, 25, 35, 39] approaches have been developed, in order to remove scene artifacts such as specularities and shadows. In contrast, Haussecker *et al.* attempted to explicitly model brightness changes during estimation [15]. In section 5.2 we show that these mitigation techniques can be complementary to the more robust ICFs.

The remainder of this paper is structured as follows. In section 3 the properties of a number of previously proposed motion estimation metrics are examined. Based on these observations, section 4 introduces and then analyses ICFs, exploiting machine learning techniques. Finally section 5 examines how the use of these robust metrics affects the performance of existing motion estimation techniques.

3 Matching Metrics

In order to discuss a range of different motion estimation metrics in the same framework, some notation must be introduced. Assuming that a collection of M sensors produces image streams $\boldsymbol{I}_{1..M}$, the pair of pixel positions from one camera associated with a particular motion vector, is defined as Γ_m. If each sensor has an associated projection function $\boldsymbol{\Pi}_{1...M}$ then the pair of pixel positions from sensor m, which support a 3D motion vector $\boldsymbol{v} = (u, v, w)$ at 3D location $\boldsymbol{r} = (x, y, z)$, is defined as

$$\Gamma_m = \left\{ \left(\boldsymbol{\Pi}_m \left(\boldsymbol{r} \right), t \right), \left(\boldsymbol{\Pi}_m \left(\boldsymbol{r} + \boldsymbol{v} \right), t + 1 \right) \right\}. \tag{1}$$

In other words, the pixels supporting the motion, are those obtained by projecting its start point to every sensor at time t, and by projecting its end point to every sensor at time $t+1$ as shown in figure 1.

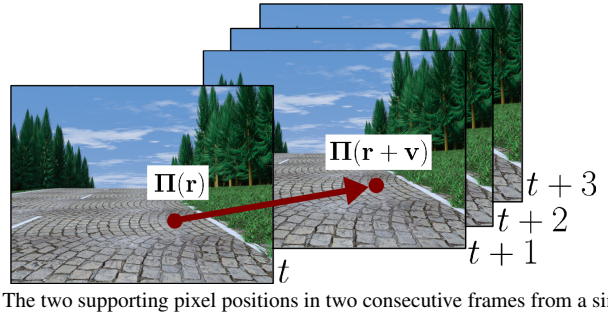


Figure 1: The two supporting pixel positions in two consecutive frames from a single camera.

Given the pairs of supporting pixel *locations* $\Gamma_{1..m}$ from each sensor, the set of associated pixel *values* Φ is obtained by indexing the relevant images

$$\Phi = \left\{ \boldsymbol{I}_m(\boldsymbol{l}) \mid \boldsymbol{l} \in \Gamma_m \forall m \right\}. \tag{2}$$

The sets of supporting pixel locations (Γ) and values (Φ), allow us to begin defining the metrics under consideration. The simplest metric used to determine the validity of a motion, is to calculate the deviation from brightness constancy according to some norm. Many such approaches exist, using the l_1 norm [27], the l_2 norm [4, 8, 17, 18, 26, 35] or their various

robust approximations such as the Charbonnier function [34, 39]. In our experiments we found that the specifics of the norm make little difference. As such, for conciseness we represent this class of cost functions in the paper as

$$SQ(\Phi) = \sum_{c \in \Phi} |c - \bar{c}|^2, \tag{3}$$

where $\bar{c}$ is the mean value of the supporting pixels (Φ). This is the equivalent of calculating the appearance variance across all the observations. See the supplementary material[1] for full results of all metrics using different norms.

Note that these metrics may be applied to RGB data, as well as greyscale, by creating a separate set of supporting pixel values Φ_c for each input channel. This leads to a metric based on the colour constancy assumption. In a similar vein, the input images may be replaced by gradient images, leading to a set of supporting gradient values Φ_g, and associated metric

$$SQ_g(\Phi_g) = \sum_{c_g \in \Phi_g} |c_g - \bar{c}_g|^2, \tag{4}$$

based on the Gradient Constancy Assumption [8, 17, 26].

A slightly more complex metric, which is commonly used, is the so called "optical flow constraint" *OFC* [7, 10, 21, 23, 32]. The formulation of this metric is slightly different, supporting pixels from time $t+1$ are not used directly. Instead they serve to create temporal gradient image $\boldsymbol{I}_t$ which is used in combination with the spatial gradient images $\boldsymbol{I}_x$ and $\boldsymbol{I}_y$

$$OFC(\Gamma) = \sum_{m=1}^{M} \sum_{\boldsymbol{l} \in \Gamma_m} \begin{cases} 0 & \text{if } t+1 \in \boldsymbol{l} \\ \boldsymbol{I}_t(\boldsymbol{l}) + u\boldsymbol{I}_x(\boldsymbol{l}) + v\boldsymbol{I}_y(\boldsymbol{l}) & \text{otherwise .} \end{cases} \tag{5}$$

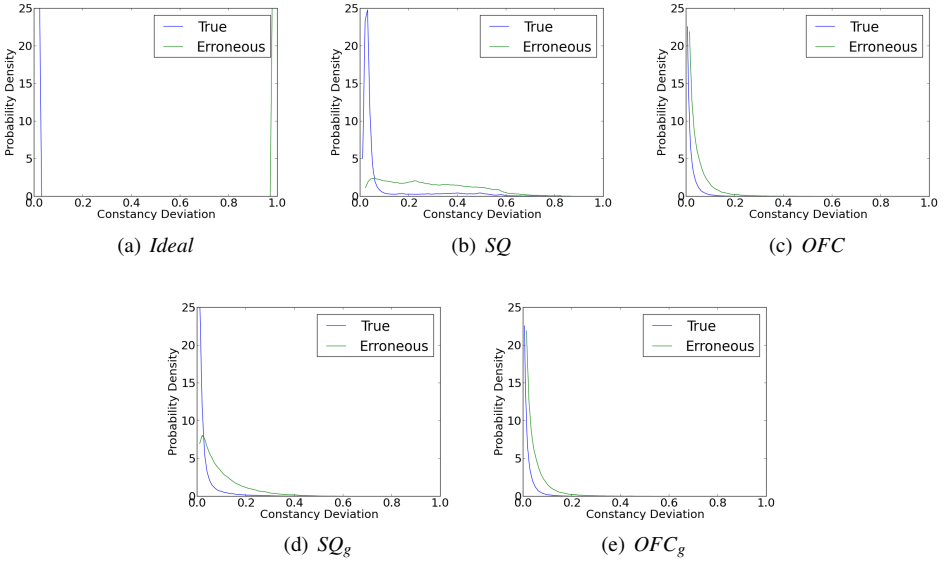
This metric equates to a linearised Taylor expansion of the brightness constancy assumption (i.e. dropping terms of quadratic or higher power).

Distinguishing truth from errors

An ideal metric should provide a low cost for true motions and a high cost for incorrect motions. Indeed, ideally the cost should continuously decay as the error decreases. Figure 2(a) shows the probability density function of responses for such an ideal metric, when applied to a true, and highly erroneous motion field. In this ideal case the true motion field registers no violation in the underlying constancy assumption (i.e. the PDF contains all responses at 0.), while the incorrect motions strongly violate the assumption, leading to a PDF concentrated at 1. Note that the responses are normalized between 0 and 1. The remainder of figure 2 illustrates the actual response distributions found for each of the previously discussed metrics, when applied to ground truth and high error motion fields from the Middlebury dataset [31]. For conciseness, this paper only presents results where the erroneous motion fields are created by motion magnitude (i.e. End-Point) errors. Preliminary testing indicated that the metrics exhibit similar behaviour under directional errors.

These results tell an unfortunate story. Most ground truth motions are assigned to the lower 20% of the responses, with the occlusion and specularity effects seen previously being the minority. However, similar responses are produced, even for the significantly erroneous motions. Indeed the linearised brightness constancy metric *OFC* shows an 80% overlap

[1] personal.ee.surrey.ac.uk/Personal/S.Hadfield/icf.html



(a) *Ideal* (b) *SQ* (c) *OFC*

(d) SQ_g (e) OFC_g

Figure 2: Left: an ideal response PDF for ground truth and error motion fields. Right: the actual distribution of responses for various motion estimation metrics, applied to the ground truth and error motion fields of a real scene. Responses are normalized in the range 0 to 1.

between the two PDFs. For the gradient based metric SQ_g the erroneous motions distribution is a heavier tailed version of the ground truth distribution. The best separation occurs for the simple brightness constancy metric SQ, however performance is still quite poor, given the large error under consideration. The overlap in the response distributions tells us that most erroneous motions are indistinguishable, from cases where the scene does not obey the constancy assumption (due to specularities, directional lighting etc). Thus, attempting to minimize the metric response across the scene, results in almost as many correct motions being discarded, as incorrect.

As mentioned earlier, we would ideally like the metric to provide a smoothly increasing cost as the amount of error in the estimated motion increases. In figure 3 we show the response of the metrics averaged over the whole scene as the amount of motion error is varied. In other words the graphs illustrate how the center of mass of the PDFs from figure 2 change as we gradually move from the Erroneous to True motion field. On the x axis, a position of 1 relates to the True motion field, while 0 relates to sever underestimation (i.e. no motion), and 2 to severe over-estimation. It is useful to examine this behaviour, as the optimization schemes used during motion estimation often rely on the gradient of the metric response in order to locate minima (assuming that this also relates to a reduction in the motion error).

The brightness constancy scheme SQ does display a general trend of reduced violation towards the true motion, despite the significant overlap seen previously. In contrast, the linearised brightness constancy constraint OFC always favours smaller motions, and on average little tendency towards the correct motion. This demonstrates why multi-scale approaches are so commonly employed, as an attempt to coerce the metric, into allowing larger flows. The Gradient constancy metric SQ_g also performs poorly, with all motions more than 10% different to the ground truth, producing roughly the same response. Thus, when such a metric is employed for motion estimation, convergence occurs extremely slowly, if at all, unless

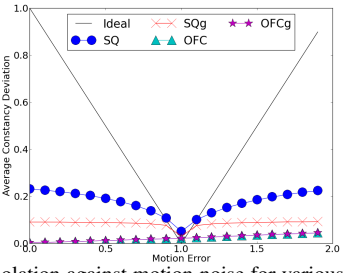


Figure 3: A plot of assumption violation against motion noise for various metrics, including an "Ideal" metric. Violation is averaged across the scene, while varying the motion error from underestimated (0) to overestimated (2).

the initialization is very close to the true value.

4 Intelligent Cost Functions

The previously discussed motion estimation metrics, propose either a simple linear or quadratic relationship, between the deviation from appearance constancy, and the "quality" of the match. However, it has been shown that many deviations occur due to the properties of the scene, without necessarily reflecting errors in the input motion. These metrics do not take into account any of these "acceptable" inconsistencies, and are thus unable to differentiate motion errors from scene artifacts. To address this issue, the use of machine learning techniques is proposed, to find an intelligent matching criteria, of unconstrained form, which is robust to the appearance inconsistencies of real data, and sensitive to inconsistencies due to motion errors. This matching criteria is represented as $\rho(\boldsymbol{\nu}|F)$, a function of the input motion given some features F extracted from the supporting pixels Γ.

Learning such a nonlinear function, allows the metric to embody more complex behaviours. As an example, it may be expected that in very light or dark parts of the scene, image contrast would be reduced. In this case, little variation may be expected naturally, and any appearance deviations may be more significant. Such behaviour would serve to flatten the exaggerated responses observed in underexposed regions. Alternatively, specular effects may cause a large change in appearance across all colour channels, while a change in appearance for only one channel is more likely to relate to an erroneous motion.

In this paper, a Gaussian Process (GP) [28] is employed to model the relationship between the input features and the level of motion error. The GP provides a non-parametric means of fitting complex data, estimating a distribution across the infinite set of possible cost functions. For all analysis in this and the following section, the GP model used an exponential kernel and was trained using 500,000 samples randomly extracted from a training set of scene flow sequences [31], with testing performed on unseen sequences.

A range of different approaches to encoding visual information into the features F were explored. The simple "baseline" approach referred to as F_{var} contains only a single element, which is equivalent to the *output* of the squared differences metric SQ from equation (3) (i.e. the variance of the appearance)

$$F_{var}(\Phi) = \left\{ \frac{1}{|\Phi|} \sum_{c \in \Phi} |c - \bar{c}|^2 \right\}. \tag{6}$$

The second approach is to take the distance of each pixel value from the mean value,

$$F_{dif}(\Phi) = \{ |c - \bar{c}| : c \in \Phi \}, \tag{7}$$

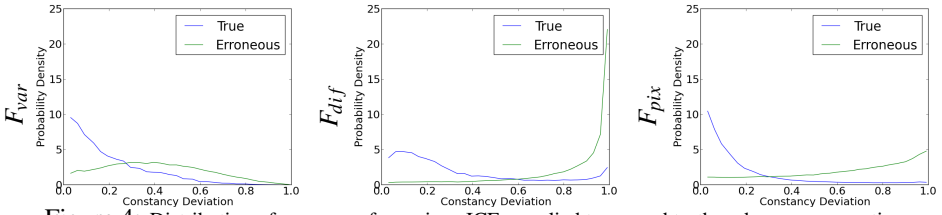


Figure 4: Distribution of responses for various ICFs, applied to ground truth and erroneous motions.

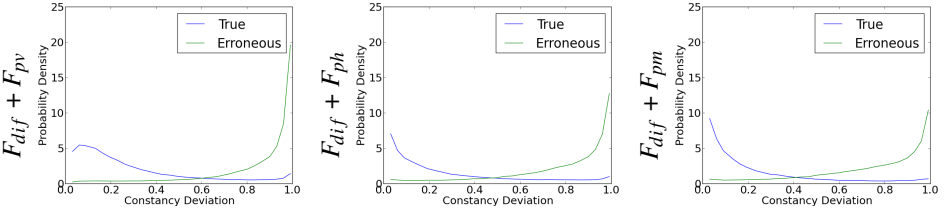


Figure 5: The distribution of responses for ICFs including contextual information.

which is equivalent to the *input* squared difference metric SQ. The third possibility considered, is allowing the ICF to learn directly from the raw pixel values

$$F_{pix}(\Phi) = \Phi. \tag{8}$$

Figure 4 shows the performance of ICFs, based on these various input encodings. Again, these results are also provided for a range of additional sequences as supplementary material, demonstrating the generality of these findings across different types of scene. Note that supplementary material results on KITTI [12] sequences are trained on greyscale versions of the Middlebury sequences [31], showing generality across significantly different domains.

Using the F_{var} features (i.e. learning a nonlinear mapping of the SQ metric), little additional separation is obtained between the classes. However, in the case of F_{dif} encoding, the ICF is able to exploit richer input features to greatly improve separation. This is due to it's ability to consider nonlinear combinations of inputs, rather than a simple remapping.

The raw F_{pix} encoding doesn't allow ICFs to distinguish motions as well as F_{dif} despite theoretically including richer information. This is perhaps unsurprising, as the machine learning within the ICF assumes independence of features, while in F_{pix}, most information is contained within the correlation between features. To illustrate this point, note that for a true motion, every difference feature F_{dif} should be low, and may be examined in isolation. However, for a true motion, the pixel features F_{pix} may take any value, as long as all features relating to one colour channel are the same.

All 3 encoding schemes lead to learned metrics which display a reduction in the effect of outliers. In the original metrics of section 3, the vast majority of motions (both true and erroneous) fell within the bottom 20% of the response range, while the remaining 80% was populated by a very small number of outliers. Even the F_{var} based ICF offers this outlier reduction, despite having little effect on class separation.

Local context

In addition to allowing complex nonlinear combinations of the information from supporting pixels, it is trivial to include higher level information in an ICF. As an example, features

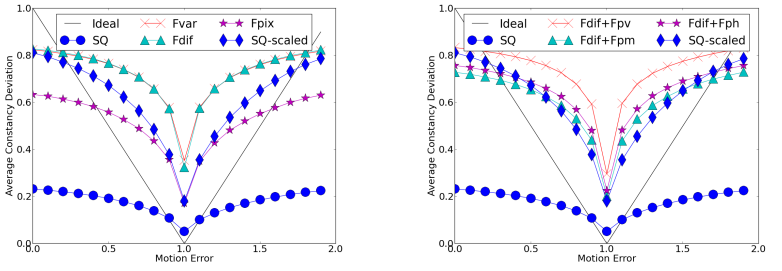


Figure 6: Average response of ICFs with and without context information, against varying levels of motion error. The best standard metric (*SQ*) is shown for comparison. Also shown is a rescaled version of *SQ* having a similar area under curve (equivalent of ignoring the top 75% of outliers).

based on local context may be included, such as the local image contrast around each supporting pixel (F_{pv}), which may be useful to distinguish motions at boundaries from motions within objects. The local mean (F_{pm}) may also encode useful information about the lighting around the supporting pixels, equivalent to a more robust version of the information present in F_{pix}. We also evaluate (F_{ph}) a coarse histogram of local intensities around each supporting pixel. This encompasses both the other contextual features, and some additional information.

For our experiments we use a 7×7 window for local context features. Note that all these contextual features are more general than direct "patch matching" techniques employed in local motion estimation algorithms. The contextual information encoded here avoids spatial information (i.e. no one-to-one correspondence is assumed between the pixels in the context region). However, ICFs may also be employed within such schemes, by replacing the pixel-wise comparison measure.

As can be seen in figure 5, the inclusion of the local variance feature F_{pv} provides little benefit. This implies that knowing if the motion is on an object boundary is not useful when determining its validity. The simplest contextual features F_{pm}, which are also the fastest to compute, actually proves to be the most valuable. The peak of the "true motion" PDF is closer to 0 deviation than for any other metric, while maintaining a similar overlap region. The more complex local histogram features F_{ph} actually prove slightly worse than both the simpler methods for contextual encoding. It is likely that the larger number of features make it difficult to determine the optimal costing function.

Figure 6 shows the average response of the learned metrics, across the scene, for varying levels of motion error. The slope of the response is far greater than even the original brightness constancy metrics. This suggests that ICFs will lead to much more rapid convergence in optimization based motion estimation, in addition to the expected gains in accuracy. We also show a rescaled version of the best brightness constancy metric, to bring the area under the curve into the same range as the ICFs. This is the equivalent of dropping the top 75% of outliers from the *SQ* metric (points with exceptionally high response), and using only the lower regions of the response PDF for calculating the center of mass. We can see that this artificially rescaled metric has a slope somewhat better than the standard ICFs, and roughly the same as the context based ICFs. This re-affirms our earlier observation that ICFs provide excellent robustness to outliers.

5 Motion Estimation with ICFs

The previous analysis has shown that standard motion estimation cost functions have some significant flaws on real data, and that greater robustness may be obtained via ICFs. However,

most motion estimation techniques already contain mechanisms such as multiscale estimation and iterative warping, designed to mitigate the inadequacies of standard cost functions. As such, it is important to examine whether the use of ICFs does in fact translate to more accurate motion estimates.

5.1 Scene Flow Estimation with ICFs

To this end, a recent, publicly available, algorithm for scene flow estimation [14] (based on the *SQ* cost function) is modified to exploit ICFs. For consistency, the resulting system is evaluated using the procedure of [14], averaging performance across all pixels (including the occlusion mask) in terms of the angular error (ε_{ae}), the within plane motion error (ε_{of}), the out of plane motion error (ε_{sf}), and the structural reconstruction error (ε_{st}). Results below are averaged over all sequences from the Middlebury dataset in [14].

The results display similar characteristics to those observed earlier in the paper. This demonstrates that there are significant gains to be made by employing ICFs, even in existing techniques which already account for the limitations of their cost function. F_{var} offers marginal improvement over the original formulation, confirming that a simple nonlinear mapping is insufficient for this more complex task. The other ICFs all provide some degree of performance gain, with the difference features providing the best individual performance. Contextual information also seems to help with $F_{dif} + F_{pm}$ providing a 44% improvement in magnitude accuracy, and 20% improvement in directional accuracy, coupled with a 30% reduction in structural error.

The runtime using each of the metrics is also listed. The additional cost of querying the ICFs proves inconsequential, and the simple feature encoding schemes such as F_{pix} actually prove faster than the original formulation.

Metric	Mode	ε_{of}	ε_{sf}	ε_{st}	ε_{ae}	Runtime (secs)
SQ [14]	Multiview	0.173	0.010	1.52	1.66	352
F_{var}	Multiview	0.164	0.021	1.53	1.63	389
F_{dif}	Multiview	0.111	0.009	**1.04**	1.41	363
F_{pix}	Multiview	0.142	0.012	1.17	1.47	**340**
$F_{dif} + F_{pv}$	Multiview	0.100	**0.005**	1.10	1.50	440
$F_{dif} + F_{ph}$	Multiview	0.134	0.008	1.14	1.59	560
$F_{dif} + F_{pm}$	Multiview	**0.098**	0.014	1.06	**1.23**	430

Table 1: Performance for scene flow estimation [14], based on the original *SQ* metric, and a range of ICFs. Also shown is the runtime for a single frame estimation, using each metric.

5.2 Optical Flow Estimation with ICFs

The potential of intelligent metrics is not limited to scene flow estimation. To demonstrate this, we also integrate specially trained ICFs into the non-local optical flow approach of Sun *et al*. [34] (the only technique in the Middlebury benchmark to provide source code rather than binaries). Note that in the case of optical flow, each motion vector has only 2 supporting pixels, meaning very few features for the ICFs to exploit. Table 5.2 compares the ICF results against the OFC_g metric originally used in [34]. Errors in this case are measured using the standard end-point-error (EPE) and angular error (ε_{ae}) defined in [3]. Again, results on additional sequences are supplied as supplementary material. The strength of ICFs is their ability to learn nonlinear relationships, however this also precludes the use of linear solvers as in [34]. To deal with this issue, without using expensive nonlinear optimisation, we instead optimise a linear Taylor approximation of the ICF as used in standard (OFC) systems.

Metric	ε_{ae}	EPE	Runtime
OFC_g [34]	3.86	0.096	**420 secs**
F_{var}	3.10	0.072	649 secs
F_{dif}	3.02	0.071	672 secs
F_{pix}	**2.87**	0.053	660 secs
$F_{dif} + F_{pv}$	3.12	0.064	757 secs
$F_{dif} + F_{ph}$	2.99	0.063	958 secs
$F_{dif} + F_{pm}$	3.01	**0.051**	732 secs

Table 2: Performance for optical flow estimation, using the OFC_g based approach of Sun *et al*. [34], and a range of ICFs.

OFC_g F_{pix} Ground truth

Figure 7: Example motion fields for one of the Middlebury sequences, comparing the original approach and an ICF against the ground truth.

The results of the optical flow experiments are similar, albeit with more modest gains. This is because, although ICFs provide a more robust cost function, there is no need to generalize across the responses characteristics of multiple cameras. The simpler nature of the task is also evident in the fact that the raw pixel features F_{pix} prove to be the most effective encoding scheme, followed by the contextual approaches. However, the originally used OFC_g metric still proves the fastest to compute, at the cost of reduced accuracy.

Figure 7 shows estimated flow fields, note that the ICFs lead to patchier results due to local minima. However, motion is better recovered in occluded regions, such as behind the shell and regions in shadow such as the center of the D and the square background cutouts.

It is interesting to note that the original technique includes explicit modelling of occlusions. However, the robustness of ICFs still brings significant performance gains, likely due to other types of scene artifact. This raises an interesting possibility; when using ICFs within a particular algorithm, it may prove valuable to employ bootstrapping techniques during training. The ICFs may then be adapted to focus learning into areas which prove problematic for the technique in question.

6 Conclusions

In conclusion, an extensive analysis has been performed for various motion estimation metrics in scene flow estimation. It has been shown that all previous metrics produce similar response distributions, for both true and erroneous motions (i.e. the underlying constancy assumption is almost as valid for incorrect motions, as it is for true motions). Motivated by these observations, "Intelligent Cost Functions" were proposed, making use of machine learning techniques. This was shown to provide a marked improvement in the separation of true and erroneous motions, meaning a high response is far more likely to indicate a motion error, than a scene artifact such as a reflection. It was also shown that this translates to improvements of 44% and 20% within existing scene flow and optical flow techniques.

As future work, the idea of ICFs may be extended to the smoothness term present in many energy surfaces. This could lead to a joint framework, which embodies both the data matching and smoothness behaviours of real data. In addition more complex patch based metrics such as normalized cross correlation [40] and mutual information [11] may prove valuable. It would also be interesting to examine if certain ICFs are better suited to particular scenarios (e.g. high noise or strong differences between sensors). A higher level system could then be used to determine which ICFs is most suitable for the current conditions.

Acknowledgements

This work was supported by the EPSRC project "Learning to Recognise Dynamic Visual Content from Broadcast Footage" (EP/I011811/1).

References

[1] Padmanabhan Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989.

[2] J.F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition modeling, algorithms, and parameter selection. *IJCV*, 67(1):111–136, 2006.

[3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[4] T. Basha, Y. Moses, and N. Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *Proc. CVPR*, pages 1506–1513, San Francisco, CA, USA, June 13 – 18 2010. doi: 10.1109/CVPR.2010.5539791.

[5] T. Basha, S. Avidan, A. Hornung, and W. Matusik. Structure and motion from scene registration. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1426 –1433, june 2012. doi: 10.1109/CVPR.2012.6247830.

[6] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *International Journal of Computer Vision*, pages 1–16, 2012. ISSN 0920-5691. doi: 10.1007/s11263-012-0542-7.

[7] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pages 231–236. IEEE, 1993.

[8] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*, pages 25–36, Prague, Czech Republic, May 11 – 14 2004. Springer.

[9] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *IJCV*, 61(3), 2005.

[10] F. Devernay, D. Mateus, and M. Guilbert. Multi-camera scene flow by tracking 3D points and surfels. In *Proc. CVPR*, volume 2, pages 2203–2212, New York, NY, USA, June 2006. doi: 10.1109/CVPR.2006.194.

[11] Nicholas Dowson and Richard Bowden. Mutual information for Lucas-Kanade tracking (MILK): An inverse compositional formulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(1):180–185, 2008.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[13] Simon Hadfield and Richard Bowden. Go with the flow: Hand trajectories in 3D via clustered scene flow. In *In Proceedings, International Conference on Image Analysis and Recognition*, Aveiro, Portugal, June25 - 27 2012. Springer.

[14] Simon Hadfield and Richard Bowden. Scene particles: Unregularized particle based scene flow estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36 (3):564–576, March 2014. doi: 10.1109/TPAMI.2013.162.

[15] Horst W. Haussecker and David J. Fleet. Computing optical flow with physical models of brightness variation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):661–673, 2001.

[16] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial intelligence*, 17 (1):185–203, 1981.

[17] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. ICCV*, pages 1–7, Rio de Janario, Brazil, October 16 – 19 2007. doi: 10.1109/ICCV.2007.4409000.

[18] M. Isard and J. MacCormick. Dense motion and disparity estimation via loopy belief propagation. In *Proc. ACCV*, pages 32–41, Hyderabad, India, January13-16 2006. Springer.

[19] Tae Hyun Kim, Hee Seok Lee, and Kyoung Mu Lee. Optical flow via locally adaptive fusion of complementary data costs. In *Proc. ICCV*, Sydney, Australia, December 3 – 6 2013.

[20] R. Li and S. Sclaroff. Multi-scale 3D scene flow from binocular stereo sequences. *CVIU*, 110(1):75–90, 2008.

[21] Rui Li and Stan Sclaroff. Multi-scale 3D scene flow from binocular stereo sequences. In *Proc. IEEE Workshop Motion and Video Computing WACV/MOTIONS '05 Volume 2*, volume 2, pages 147–153, 2005. doi: 10.1109/ACVMOT.2005.80.

[22] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.

[23] T.C. Lukins and R.B. Fisher. Colour constrained 4D flow. In *Proc. BMVC*, pages 340–348, Oxford, UK, September 6 – 8 2005.

[24] Oisin Mac Aodha, Ahmad Humayun, Marc Pollefeys, and Gabriel J Brostow. Learning a confidence measure for optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1107–1120, 2013.

[25] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, volume 22. Amer Mathematical Society, 2001.

[26] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.

[27] Clemens Rabe, Thomas Müller, Andreas Wedel, and Uwe Franke. Dense, robust, and accurate motion field estimation from stereo image sequences in real-time. In *Proc. ECCV*, Heraklion, Crete, September 5 – 11 2010.

[28] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[29] J. Ruttle, M. Manzke, and R. Dahyot. Estimating 3D scene flow from multiple 2D optical flows. In *Proc. International Machine Vision and Image Processing Conference*, pages 1–6, Dublin, Ireland, September2-4 2009. doi: 10.1109/IMVIP.2009.8.

[30] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 80(1):72–91, 2008.

[31] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, volume 1, 2003. doi: 10.1109/CVPR.2003.1211354.

[32] T. Schuchert, T. Aach, and H. Scharr. Range flow for varying illumination. In *Proc. ECCV*, pages 509–522, Marseille, France, October 12 – 18 2008. IEEE.

[33] D. Sun, S. Roth, J. Lewis, and M. Black. Learning optical flow. In *Proc. ECCV*, pages 83–97, Marseille, France, October 12 – 18 2008. IEEE, Springer.

[34] Deqing Sun, S. Roth, and M.J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432 –2439, june 2010. doi: 10.1109/CVPR.2010.5539939.

[35] W. Trobin, T. Pock, D. Cremers, and H. Bischof. An unbiased second-order prior for high-accuracy motion estimation. *Pattern Recognition*, pages 396–405, 2008.

[36] Levi Valgaerts, Andres Bruhn, Henning Zimmer, Joachim Weickert, Carsten Stoll, and Christian Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In *Proc. ECCV*, 2010.

[37] S. Vedula, S. Baker, S. Seitz, and T. Kanade. Shape and motion carving in 6D. In *Proc. CVPR*, volume 2, pages 592–598, Hilton Head, SC, USA, June 13 – 15 2000. doi: 10.1109/CVPR.2000.854926.

[38] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Proc. ECCV*, volume 5302, pages 739–751, Marseille, France, October 12 – 18 2008. IEEE, Springer, Heidelberg.

[39] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for TV-L1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, pages 23–45, 2009.

[40] Manuel Werlberger, Thomas Pock, and Horst Bischof. Motion estimation with non-local total variation regularization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2464–2471, 2010.

[41] Li Xu, Jiaya Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1744 –1757, sept. 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.236.

[42] Ye Zhang and Chandra Kambhamettu. On 3-D scene flow and structure recovery from multiview image sequences. *IEEE Transactions on Systems, Man and Cybernetics*, 33: 592–606, 2003. doi: 10.1109/TSMCB.2003.814284.