Improved Depth Recovery In Consumer Depth Cameras via Disparity Space Fusion within Cross-spectral Stereo

Gregoire Payen de La Garanderie
gregoire@hochet.info

School of Engineering
Cranfield University
Bedfordshire, UK

Toby P. Breckon
toby.breckon@durham.ac.uk

School of Engineering and Computing Sciences, Durham University
Durham, UK

Abstract

We address the issue of improving depth coverage in consumer depth cameras based on the combined use of cross-spectral stereo and near infrared structured light sensing. Specifically we show that fusion of disparity over these modalities prior to subsequent optimization, within the disparity space image, facilitates the recovery of scene depth information in regions where structured light sensing alone fails. This joint approach, leveraging disparity information from both structured light and cross-spectral stereo, facilitates the recovery of global scene depth comprising both texture-less object depth, where stereo sensing commonly fails, and highly reflective object depth, where structured light active sensing commonly fails. The proposed solution is illustrated using dense gradient feature matching and is shown to outperform prior approaches that use late-stage fused cross-spectral stereo depth as a facet of improved sensing for consumer depth cameras.

1 Introduction

Low-cost consumer depth cameras have risen to widespread prevalence across many areas of 3D computer vision [19]. This has seen the combined use of colour and 3D depth most commonly leverage the use of near infrared structured light projection (830nm) with regular visible-band colour sensing (400-700nm) to provide co-registered colour (RGB) and depth (D) as combined RGB-D image components. In terms of real-time performance and general depth accuracy, this outperforms common stereo vision approaches on texture-less object surfaces [10] but notably fails in depth recovery for transparent, specular and reflective surface regions [2].

Based on the common physical characteristics of such devices - comprising a colour camera, an infrared pattern projector and corresponding infrared camera (e.g. Microsoft Kinect / PrimeSense Carmine, [19]) - an obvious yet commonly under-utilized cross-modal stereo configuration exists between the two camera sensors.

Prior work on this topic (Chiu *et al.* [2]) has already considered such a cross-spectral formulation to improve depth recovery. This work centres on the recovery of an optimized pseudo-infrared version of the colour image, I_{ir}^{pseudo}, via a weighted combination of the individual RGB colour channels [2], such that a conventional stereo matching approach can then be applied in conjunction with the sensed infrared image [9] over the resulting $\{I_{ir}^{pseudo}, I_{ir}\}$ stereo image pair. An alternative scheme based on using any one or all of three colour channels to form a stereo pair, $\{I_c, I_{ir}\}$ for $c \in \{R, G, B\}$, is also considered but is outperformed by the former approach [2]. Later work [3] replaces the scalar weights with localized 3×3 patch filters derived from prior learning but achieves only a marginal performance improvement. In both cases, a simple late-stage union of the resulting stereo depth information with that from structured light shows improved depth recovery for transparent and specular surfaces where conventional consumer depth cameras of this configuration fail [2, 3].

Related work has considered the topic of fusing time-of-flight (ToF) camera sensing with stereo depth [23]. Zhu *et al.* [23] use global MAP-MRF optimization to solve the stereo vision problem via posteriors based on ToF depth within the optimization phase. Later work refined this by introducing the notion of reliability of each depth image within the fusion framework [24].

Other work has considered depth recovery improvement in depth cameras as a classical image in-painting problem [17]. Qi *et al.* [17] successfully tackles the in-painting of depth shadows due to object occlusion [1] without considering the challenges transparent and specular surfaces. Alternatively [21] augment the existing RGB-D sensing set-up with an additional colour stereo sensor to show improved depth recovery on human faces following fusion method of [23].

Work on the conventional cross-spectral stereo problem, outside of the specifics of RGB-D depth recovery improvement, is more developed [12, 13, 15, 20]. Krotosky and Trivedi [12, 13] investigate cross-spectral stereo for pedestrian detection and tracking, using a window-based Mutual Information (MI) approach inspired by the original work of Egnal [5]. However, depth computation is only performed for isolated objects (i.e. pedestrians) via prior foreground extraction and subsequent localised stereo matching [12, 13]. Krotosky and Trivedi [13] additionally demonstrate the failure of dense depth computation using MI in the global energy minimisation framework of [8] caused by the lack of a global intensity transform between the images. Torabi and Bilodeau [20] describe a very similar window-based approach but replace MI by Local Self-Similarity (LSS) as a correspondence measure. LSS was originally proposed in [18] for object detection, retrieval and action recognition in visually differing scenes and better performance than MI for this task is reported but again only on isolated scene objects [20]. More recently, Pinggera *et al.* [15] introduce a new dense depth recovery approach using Histogram of Oriented Gradient (HOG) feature descriptors [4] to address the problem of cross-spectral stereo matching between colour and far-infrared images. Essentially HOG descriptors are computed for each pixel to capture local geometry, invariant to spectral image characteristics, and are shown in [15] to notably outperform prior work [13, 20] and conventional radiometric (illumination) invariant stereo matching techniques [10].

With reference to this prior work on cross-spectral stereo, the previously discussed pseudo-infrared driven depth improvement approach proposed by Chiu et al. [2, 3] most closely follows the early work of a number of authors [5, 6, 7, 11] using simulated cross-spectral data (i.e. where one image has undergone a radiometric transform to simulate an infrared image). Recently Pinggera *et al.* [15] showed the limited applicability of such approaches [5, 6, 7, 11] in comparison to the use of dense gradient features with an appropriate optimisation approach.

By contrast to this earlier RGB-D depth improvement prior work [2, 3], we propose the use of *"best in class"* dense gradient features from [15] to facilitate recovery of secondary cross-spectral stereo disparity directly from the depth camera $\{I_{ir}, I_{RGB}\}$ image pair. Our main contribution is the fusion of this secondary cross-spectral (CS) disparity information with *a priori* depth information, obtained via conventional structured light (SL) sensing, within the disparity space image constructed prior to conventional disparity optimization for scene depth recovery. Our work extends that of [2], which focused its evaluation on boundary recovery for object segmentation, to provide dense depth recovery spanning object occlusions in addition to transparent and specular objects. This provides combined textured and texture-less surface depth recovery from a single consumer depth camera without the need for additional sensor augmentation or adaptation [23].

2 Proposed Approach

We consider a cross-spectral (CS) stereo matching process, based on [15], into which we fuse *a priori* depth information obtained from structured light (SL). Our discussion focuses on the recovery of disparity (calibrated pixel-wise differences) from which scene depth is subsequently recovered based on established stereo calibration techniques [22] (using a calibration target visible in both spectral bands [15]). Our subsequent approach can be split into three steps:- 1) match cost computation, 2) disparity optimisation and 3) disparity fusion.

2.1 Match Cost Computation

We consider the stereo matching cost computation approaches proposed by [2] and [15]. Chiu *et al.* [2] propose the construction of a single channel pseudo-infrared image from the individual RGB colour components. This is constructed via optimization to recover a set of weights $\{w_r,\ w_g,\ w_b\}$ corresponding to each of the $\{R, G, B\}$ colour components that maximizes the number of stereo disparity matches for the chosen disparity optimization technique (Eqn. 1).

$$\begin{aligned} \max_{w_r, w_g, w_b} \quad & \#\ \text{stereo matches} \{ I_{ir}^{pseudo} = w_r I_r + w_g I_g + w_b I_b, I_{ir} \} \\ \text{subject to} \quad & w_r + w_g + w_b = 1 \end{aligned} \tag{1}$$

Based on this formulation, the matching cost, $C_{ir}^{pseudo}(x, y, d)$, is then computed directly as the pixel difference between this pseudo-infrared image, I_{ir}^{pseudo}, and the true infrared image, I_{ir}, obtained from the depth camera itself for each pixel location (x, y) and stereo disparity, d. The optimization problem, to recover the required set of colour component weights $\{w_r, w_g, w_b\}$, is simply solved using grid search over the plane $w_r + w_g + w_b = 1$. However, no direct relation exists between the colour and infrared pixel values, as colour is visible-band illumination dependant whilst infrared is jointly dependant upon illumination and material reflectivity (in/to infrared light). Subsequently, the resulting $\{w_r, w_g, w_b\}$ will be different for each and every scene making this approach computationally demanding for any practical use. Computational saving could possibly made by employing a fixed pseudo-infrared formulation such as [5, 6, 7, 11] but this has been shown to compromise matching performance [15].

By contrast Pinggera *et al.* [15], motivated by the observation of $\{I_{ir} \leftrightarrow I_{RGB}\}$ pixel intensity dis-similarity yet localised image structure similarity, identify dense gradient features as an optimal cross-spectral stereo matching approach. As such, [15] proposes a matching cost based on a variant of the Histograms of Oriented Gradient (HOG) descriptor [4]. The HOG descriptor is based on histograms of oriented gradient responses in a local region around the pixel of interest. Here a rectangular block, pixel dimension $b \times b$, centred on the pixel of interest is divided into $n \times n$ (sub-)cells and for each cell a histogram of unsigned gradient orientation is computed (quantised into H histogram bins for each cell over the interval $(0,\pi)$). The histograms for all cells are then concatenated to represent the HOG descriptor for a given block (i.e. associated pixel location). For image gradient computation centred gradient filters $[-1,0,1]$ and $[-1,0,1]^T$ are used as per [4]. To maximise invariance the whole descriptor is normalised to the L2 unit norm with the resulting HOG descriptor as a $n \times n \times H$ description vector per pixel. A comparison, hence matching cost $C_{HOG}(x,y,d)$, between two HOG descriptors at pixel positions (x,y) and $(x+d,y)$ (assuming scan-line stereo rectification) is thus computed using the L1 distance. Dense HOG descriptors for every image pixel are computed efficiently by using integral histograms [16] allowing fast descriptor computation but preventing the use of spatial weighting (e.g. Gaussian) of gradient responses within any given descriptor in this case.

2.2 Disparity Optimization

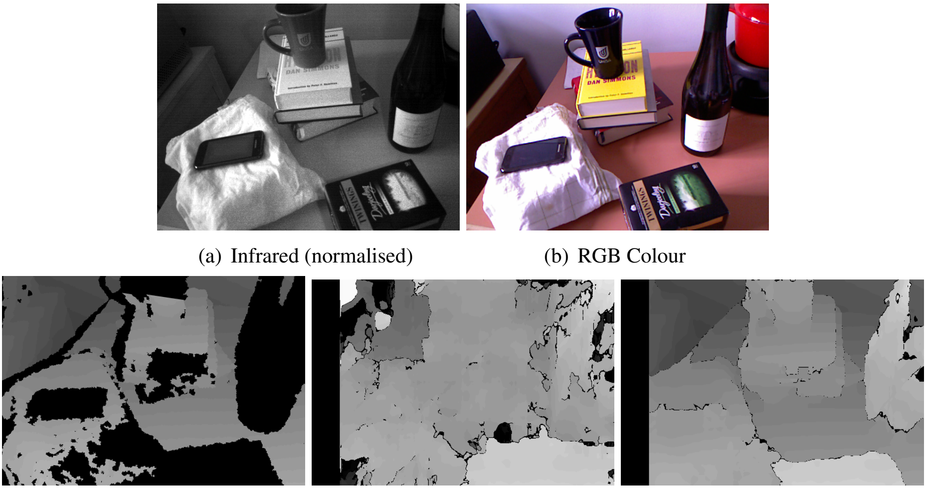
The disparity optimization step computes the disparity result $D(x,y)$ from the specified cost matching function $C(x,y,d)$. Here, following the earlier cross-spectral stereo work of [2] and [15] we similarly use Hirschmueller's seminal Semi-Global Matching (SGM) [9] which is both computationally efficient and provides improved global disparity smoothness constraints compared to alternative approaches [10, 14]. Within our SGM optimization we additionally specify a uniqueness ratio, u, such that disparity, d, corresponding minimum match cost, $\min_{c()} C(x,y,d)$, should be considered valid for pixel location (x,y) only if the next largest match cost for alternative disparity d', $C(x,y,d')$ satisfies $\frac{C(x,y,d')-C(x,y,d)}{C(x,y,d')} - u > 0$.

This allows us to control the disparity optimization in order to filter out poor quality disparity information originating from potentially ambiguous stereo matches (i.e. when the difference between minimal $C(x,y,d)$ and next possible $C(x,y,d')$ disparity solution is small). Within our depth recovery context, this allows us to filter out unreliable disparity estimates within texture-less scene regions from the resulting disparity image. Within such regions depth recovered from structured light projection will be reliable whilst conversely stereo depth will be of greater reliability within highly textured regions (including transparent and specular surfaces).

2.3 Disparity Fusion

In the prior work of [2], disparity fusion between structured light sensing and cross-spectral stereo is performed based on a simple union in projected depth space based on co-registration from *a priori* calibration [22]. Disparity, d, to depth, z, projection is carried out based on the relationship $d = \frac{fB}{z}$ using camera focal length, f, and stereo base-line, B, recovered during this earlier stereo calibration. A corresponding disparity image is then recovered via back-projection from depth space using the same formulation. In cases where this results in a *"disparity collision"* (i.e. two 3D depth space points are projected to a single 2D disparity image pixel) we favour structured light derived disparity over cross-spectral stereo. This essentially performs the conditional union of disparity from structured light (SL) and cross-spectral stereo (CS via $C_{ir}^{pseudo}(x,y,d)$) such that SL is favoured over CS when present.

By contrast, we propose using disparity fusion that is integral to the disparity optimization itself and essentially performs such a union in early-stage disparity cost space rather than in a late-stage union of the resulting disparity following [2]. This is achieved by modifying the *disparity space image*, formed by $C(x,y,d)$, which constitutes the disparity cost space over which disparity optimization will be performed. We construct an alternative cost



(a) Infrared (normalised)

(b) RGB Colour

(c) SL disparity (from device)

(d) CS disparity ($C_{HOG}(x,y,d)$)

(e) CS-DSI disparity (proposed)

Figure 1: Fused disparity estimation:- $(SL, CS) \rightarrow CS - DSI$

function, $C_{DSI}(x,y,d)$ such that the use of disparity from structured light sensing, $D_{SL}(x,y)$, is incorporated as follows:

$$C_{\mathrm{DSI}}(x,y,d) = \begin{cases} C_{HOG}(x,y,d) & \text{if } D_{SL}(x,y) \text{ is unavailable at pixel } (x,y) \\ low_c & \text{if } d = D_{SL}(x,y) \\ high_c & \text{if } d \neq D_{SL}(x,y) \end{cases} \tag{2}$$

Following from Eqn. 2, in cases where no SL disparity is available we revert to CS with its associated cost $C_{HOG}(x,y,d)$. Alternatively where there is agreement between both sensing modalities we favour this ($C_{\mathrm{DSI}}(x,y,d) = low_c$, i.e. low cost) and where we have disagreement we heavily penalize disparity case ($C_{\mathrm{DSI}}(x,y,d) = high_c$, i.e. high cost) within the overall disparity cost space.

Empirically we use $low_c = 0$ and $high_c = e^{37}$ in this work. This results in a formulation, as illustrated in Fig. 1, where an incomplete disparity result from structured light (SL, Fig. 1(c)) combined with poor-quality, otherwise largely incomprehensible cross-spectral stereo disparity (CS, Fig. 1(d)) and be successfully fused within the *disparity space image* to provide a complete and comprehensible scene disparity result that is representative of the scene objects and surfaces (CS-DSI, Fig. 1(e) compared to scene of Fig. 1(a) / 1(b)).

3 Evaluation

We present our evaluation over a number of example scenes based on comparing the disparity, as a function of depth [22], recovered using the varying approaches. Explicitly we compare the proposed disparity space image approach (CS-DSI), the original disparity obtained via structured light sensing (SL), the disparity obtained via cross-spectral stereo in isolation via $C_{HOG}(x,y,d)$ [15] (CS) and that obtained via prior cross-spectral depth improvement work of [2] (CS-union) (Figs. 1, 2, 5). This is supported by quantitative evaluation against ground truth based on analysis of the overall match quality in relation uniqueness ratio, u, specified within the SGM disparity optimization in use (Fig. 3 & 4(a) / 4(b)).

All data is collected using a Microsoft Kinect X360 depth camera providing 1280×1024 resolution $\{I_{RGB}, I_{ir}\}$ image pairs (for CS) and 640×480 resolution depth, I_{depth} (SL, upscaled for use in fusion). Static scenes are used for evaluation whereby the infrared projector used for initial depth image capture (internal to the sensor unit) is subsequently de-activated for capture of the corresponding $\{I_{RGB}, I_{ir}\}$ pair. This results in a infrared image unaffected by SL pattern projection. Our CS-DSI and CS approaches uses HOG parameters ($H = 9$, $n = 3$, $b = 18$) following [15] with SGM uniqueness ratio, u, varied in the range $\{0 \rightarrow 0.5\}$ and block-size=11.

Figure 2 shows the disparity results obtained from the $\{I_{RGB}, I_{ir}, I_{depth}\}$ triplet shown in Fig. 2(a) - 2(c) for both of the CS-union (Fig. 2(e)) and the proposed CS-DSI (Fig. 2(f)) against illustrative ground truth depth derived using manual depth labelling (Fig. 2(d)). The resulting CS-DSI disparity (Fig. 2(f)) presents a clearer disparity image with notably less missing disparity values and noise than CS-union (Fig. 2(e)) and the original SL disparity (Fig. 2(c)). The improvement in quality of the disparity resulting from CS-DSI, against the original SL disparity, is further supported by the earlier example shown in Fig. 1.

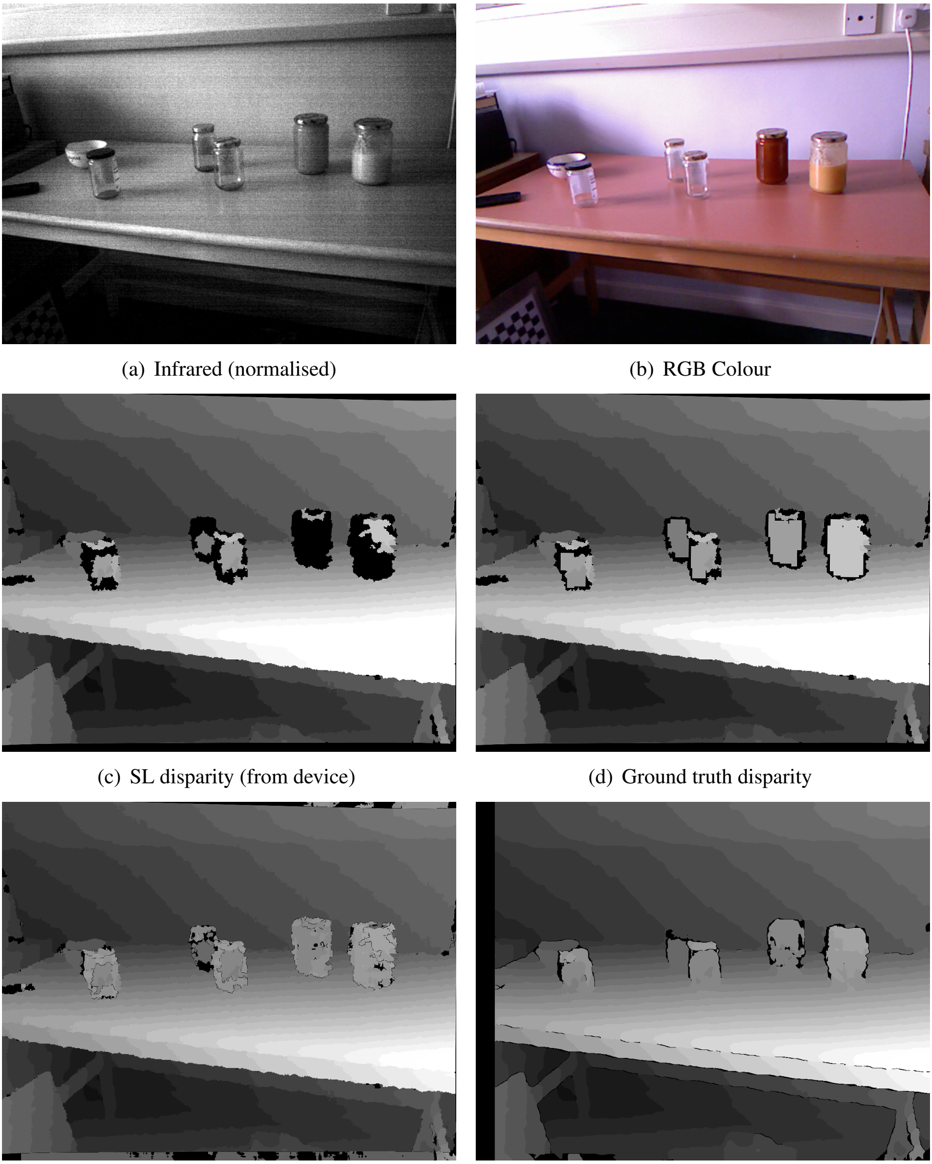
Within the CS-DSI and CS-union results shown in Fig. 2 we perform SGM disparity optimization over either the disparity space image obtained using $C_{ir}^{pseudo}(x, y, d)$ (for CS-union, as per [2]) or that obtained using $C_{\mathrm{DSI}}(x, y, d)$ which in turn uses $C_{HOG}(x, y, d)$ (as per [15]). Figure 3 shows the resulting disparity optimization using only CS in isolation (via $C_{HOG}(x, y, d)$ with SGM) for varying SGM uniqueness ratios, u, with disparity regions matching corresponding ground truth (Fig. 2(d)) shown in green. In general we note the poor performance of using CS in isolation on the texture-less scene regions (background) and consistent performance on textured regions despite transparency and specularity (foreground jars, Fig. 3).

As we can see in Fig. 3 (grey and green regions), as our uniqueness ratio, u, is increased our matching criteria gets stricter resulting in a decrease in the number of overall matches obtained. Notably we can also observe that the ratio of good matches (in green, i.e. matching ground truth) to overall matches (green + grey) increases as u is increased. Based on this observation we can introduce a ratio of good matches (against ground truth) to total matches obtained, M_{good}, and similarly a ratio of total matches obtained to all possible matches within the scene, M_{total}.

From the graph in Figure 4(a) we can observe that the ratio of good matches, M_{good}, decreases when we try to increase the overall number of matches, M_{total}, with respect to varying u for both matching costs considered here. Similarly, for both matching costs, we obtain $M_{good} > 0.8$ for higher values of u (i.e. very strict matching) but with less selective matching (lower u values) we obtain $M_{good} < 0.6$. However, $C_{HOG}(x, y, d)$ notably outperforms $C_{ir}^{pseudo}(x, y, d)$ in all cases (Fig. 4(a)) showing that our HOG matching formulation consistently outperforms the pseudo-infrared approach of [2] under varying disparity optimization conditions.

In Figure 5, we present the disparity results obtained from the $\{I_{RGB}, I_{ir}, I_{depth}\}$ triplet shown in Fig. 5(a) - 5(c). Based on the SL disparity shown in Fig. 5(c), we create an artificially challenging depth recovery scenario by removing the primary foreground objects in the SL disparity image (see Fig. 5(d)). This modified SL disparity, $D(x, y)$, is then used as the input to in our earlier formulations for CS-union (Eqn. 1) and CS-DSI (Eqn. 2) to produce the results shown in Fig. 5(e) (CS-union) and Fig. 5(f) (CS-DSI). The results show the recovery of the missing scene disparity in both cases with lesser disparity holes and greater clarity in the CS-DSI result (comparing Fig. 5(e) to Fig. 5(f)).

Notably, the performance of CS in isolation (using $C_{HOG}(x, y, d)$) is worse than CS-DSI (or CS-union) for the foreground object disparity recovery (Fig. 5(g) and 5(h)). This illustrates the additional constraint brought to the disparity recovery process via the combined



(a) Infrared (normalised)

(b) RGB Colour

(c) SL disparity (from device)

(d) Ground truth disparity

(e) CS-Union disparity [2]

(f) CS-DSI disparity (proposed)

Figure 2: Disparity recovery on transparent and specular objects

use of the remaining background SL disparity (Fig. 5(d)) within the overall disparity optimization process. Essentially we show that in the case of no local SL disparity constraint within disparity recovery, the CS-DSI method does not simply resort to a simple CS driven process in these regions (akin to [15]). As such it is not simply performing *depth filling* [1] by alternating between the SL and CS modality as required. Instead the presence of SL disparity globally constrains the overall CS-DSI disparity estimation process such that the foreground result present in Fig. 5(f) (CS-DSI) is superior to that of using CS alone (Fig. 5(g) and/or 5(h))) and indeed SL alone (Fig. 5(c)). The CS-DSI results presented in Figs. 5

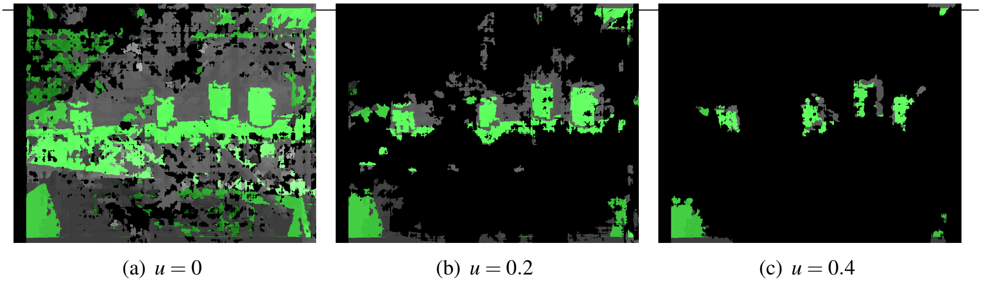
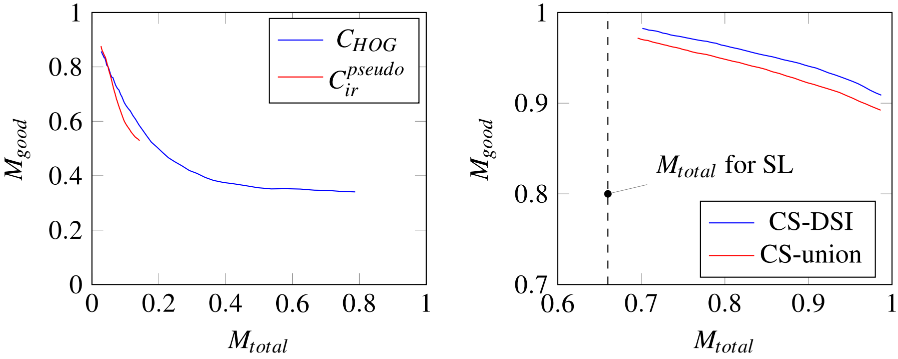


Figure 3: CS disparity image with varying SGM uniqueness ratio, u.



(a) Good match ratio vs. total matches ratio for different matching costs, $u = \{0 \rightarrow 0.4\}$

(b) Good match ratio vs. total matches ratio for different disparity fusion approaches, $u = \{0 \rightarrow 0.5\}$.

Figure 4: Match ratio comparisons for matching cost and disparity fusion approaches

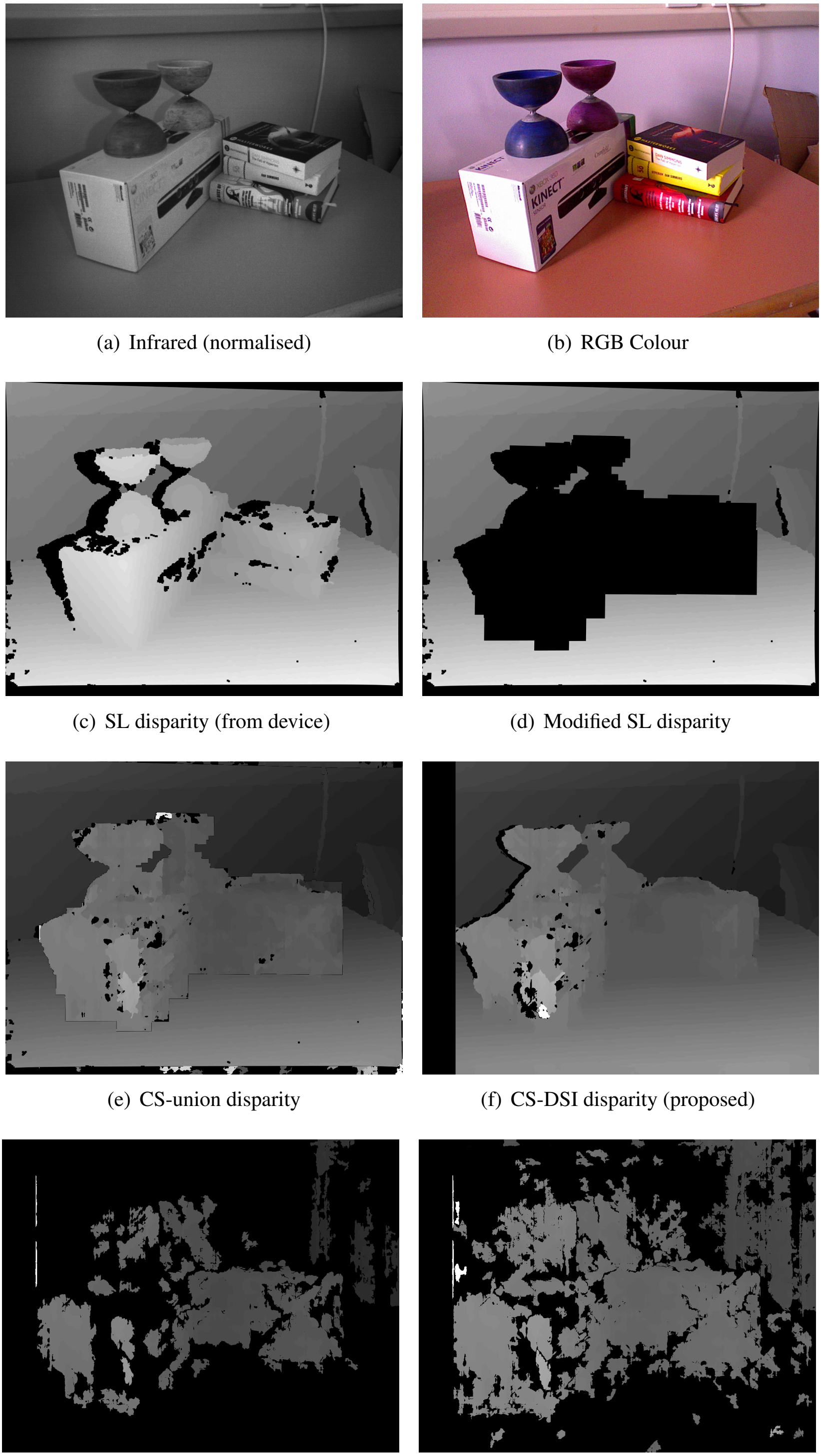
and 1 support the case that this fused disparity output via this process is indeed *"greater than the sum of the individual parts"*. By contrast, the simple depth-union based approach of [2] to which we compare (CS-union) is very much simply the sum of the individual parts.

Using our original SL disparity (Fig. 5(c)) as ground truth against our modified version (Fig. 5(d)) we can compute our earlier M_{good} and M_{total} match ratios to provide a quantitative measure of relative performance for both disparity fusion approaches (Fig. 4(b)). From Figure 4(b), we can see that the CS-DSI disparity fusion method consistently outperforms the CS-union approach [2] over a range of SGM uniqueness ratio values, u, in terms of the M_{good} and M_{total} match ratio metrics.

Practical Issues: As reported in [2] the Microsoft Kinect X360 hardware does not facilitate simultaneous capture of the RGB, $\{I_{RGB}\}$, and infrared, $\{I_{ir}\}$, video streams. Maximal speeds of 1.5-2 fps are achievable via stream switching which is still viable for many applications. De-activation of the infrared projector is currently performed manually and depth, $\{I_{depth}\}$, image capture is limited to 640×480 resolution. Within this work, these limitations are considered bespoke to this particular consumer depth camera based around its original design criteria.

4 Conclusions

Improved disparity can be recovered from a consumer depth camera based on the fusion of cross-spectral stereo and existing structured light sensing performed prior to conventional disparity space optimization within the disparity space image. Missing depth information is recovered for transparent and specular objects in addition to that missing due to inter-object occlusions and other sensing noise.



(a) Infrared (normalised)

(b) RGB Colour

(c) SL disparity (from device)

(d) Modified SL disparity

(e) CS-union disparity

(f) CS-DSI disparity (proposed)

(g) CS, C_{HOG}, $u = 0$ disparity

(h) CS, C_{HOG}, $u = 0.1$ disparity

Figure 5: Disparity recovery over large scale disparity holes

This directly extends prior work [2, 3] which is shown to produce lesser disparity recovery and requires computational expensive scene dependant optimization. By contrast, we offer improved depth recovery from a single sensing unit (consumer depth camera) without the need for individual per scene optimization, making it highly suitable for mobile sensing applications and dynamic scenes. This extends both the work of [21], which requires additional sensors to achieve the similar results, and the disparity in-painting approach of [17] which does not readily recover transparent and specular object disparity.

Future work will investigate the negation of structured light pattern effects within the infrared component of the process [3] and the generality of the proposed disparity space image fusion approach for use with alternative stereo matching and disparity optimization.

References

[1] T.P. Breckon and R.B. Fisher. 3D surface relief completion via non-parametric techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2249 – 2255, December 2008. doi: 10.1109/TPAMI.2008.153.

[2] W. C. Chiu, U. Blanke, and M. Fritz. Improving the Kinect by Cross-Modal Stereo. In *Proceedings of the British Machine Vision Conference*, pages 1–10, 2011. doi: 10.5244/C.25.116.

[3] W. C. Chiu, U. Blanke, and M. Fritz. I spy with my little eye: Learning optimal filters for cross-modal stereo under projected patterns. *Proc. IEEE International Conference on Computer Vision Workshops*, pages 1209–1214, 2011. doi: 10.1109/ICCVW.2011.6130388.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005. doi: 10.1109/CVPR.2005.177.

[5] G. Egnal. Mutual information as a stereo correspondence measure. Technical report, University of Pennsylvania, 2000.

[6] C. Fookes and S. Sridharan. Investigation & comparison of robust stereo image matching using mutual information & hierarchical prior probabilities. In *Proc. Second International Conference on Signal Processing and Communication Systems*, pages 1–10. IEEE, 2008. doi: 10.1109/ICSPCS.2008.4813750.

[7] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *Proceedings Second International Symposium on 3D Data Processing, Visualization and Transmission*, pages 961–968. IEEE, 2004. doi: 10.1109/TDPVT.2004.1335420.

[8] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 807–814. IEEE, 2005. doi: 10.1109/CVPR.2005.56.

[9] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. doi: 10.1109/TPAMI.2007.1166.

[10] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009. doi: 10.1109/TPAMI.2008.221.

[11] J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *Proc. International Conference on Computer Vision*, volume 2, pages 1033–1040. IEEE, 2003. doi: 10.1109/ICCV.2003.1238463.

[12] S. Krotosky and M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106 (2-3):270–287, 2007. doi: 10.1016/j.cviu.2006.10.008.

[13] S. Krotosky and M. Trivedi. Registering multimodal imagery with occluding objects using mutual information: application to stereo tracking of humans. In R.I. Hammoud, editor, *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*, chapter 14, pages 321–347. Springer, 2009.

[14] F. Mroz and T.P. Breckon. An empirical comparison of real-time dense stereo approaches for use in the automotive environment. *EURASIP Journal on Image and Video Processing*, 2012(13):1–19, 2012. doi: 10.1186/1687-5281-2012-13.

[15] P. Pinggera, T.P. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *Proc. British Machine Vision Conference*, pages 526.1–526.12, September 2012. doi: 10.5244/C.26.103.

[16] F. Porikli. Integral histogram: a fast way to extract histograms in Cartesian spaces. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 829–836. IEEE, 2005. doi: 10.1109/CVPR.2005.188.

[17] F. Qi, J. Han, P. Wang, G. Shi, and F. Li. Structure guided fusion for depth map inpainting. *Pattern Recognition Letters*, 34(1):70–76, January 2013. doi: 10.1016/j.patrec.2012.06.003.

[18] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. doi: 10.1109/CVPR.2007.383198.

[19] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. Computer Vision and Pattern Recognition*, pages 1297–1304. IEEE, June 2011. doi: 10.1109/CVPR.2011.5995316.

[20] A. Torabi and G.A. Bilodeau. Local self-similarity as a dense stereo correspondence measure for thermal-visible video registration. In *Proc. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 61–67, 2011. doi: 10.1109/CVPRW.2011.5981751.

[21] Y. Wang and Y. Jia. A fusion framework of stereo vision and kinect for high-quality dense depth maps. *Proc. Asian Computer Vision Conference Workshops*, pages 109–120, 2013.

[22] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. doi: 10.1109/34.888718.

[23] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587761.

[24] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan. Reliability Fusion of Time-of-Flight Depth and Stereo for High Quality Depth Maps. *IEEE transactions on pattern analysis and machine intelligence*, 33(7):1400–1414, August 2010. doi: 10.1109/TPAMI.2010.172.