On Cross-Spectral Stereo Matching using Dense Gradient Features

Peter Pinggera[1][2]
pinggera@alumni.tugraz.at
Toby Breckon[1]
toby.breckon@cranfield.ac.uk
Horst Bischof[2]
bischof@icg.tugraz.at

[1] School of Engineering, Cranfield University, Bedfordshire, UK.
[2] Institute for Computer Graphics and Vision, TU Graz, Austria.

Abstract

We address the problem of scene depth recovery within cross-spectral stereo imagery (each image sensed over a differing spectral range). We compare several robust matching techniques which are able to capture local similarities between the structure of cross-spectral images and a range of stereo optimisation techniques for the computation of valid depth estimates in this case. Specifically we deal with the recovery of dense depth information from thermal (far infrared spectrum) and optical (visible spectrum) image pairs where large differences in the characteristics of image pairs make this task significantly more challenging than the common stereo case. We show that the use of dense gradient features, based on Histograms of Oriented Gradient (HOG) descriptors, for pixel matching in combination with a strong match optimisation approach can produce largely valid, yet coarse, dense depth estimates suitable for object localisation or environment navigation. The proposed solution is compared and shown to work favourably against prior approaches based on using Mutual Information (MI) or Local Self-Similarity (LSS) descriptors.

1 Introduction

The performance of standard optical camera systems can be severely affected by environmental conditions like low lighting, shadows, smoke/dust or semi/fully camouflaged objects [9]. A method to overcome such problems is to use combined sensing systems operating in differing parts of the electromagnetic spectrum. For example, infrared images, often referred to as thermal images, are independent of visible light illumination and shadows, relatively robust to dust/smoke and can often distinguish objects which look similar to the background within the visible spectrum [13]. However, thermal images are conversely affected by ambient temperature and can offer difficulty in identifying objects with a similar temperature to the background (ambient temperature). As a result, an attractive solution is the combination of both optical and thermal images in many sensing and surveillance scenarios as the complementary nature of both modalities can be exploited and the individual drawbacks largely compensated (e.g. [27, 3, 9, 13, 7]).

Practically, a possibility is to simply alternate between the use of optical and thermal cameras - for example to switch from optical to thermal imagery for night- or low-light-vision. More sophisticated approaches use both modalities simultaneously when the circumstances permit and employ sensor fusion methods to combine the information acquired from

http://dx.doi.org/10.5244/C.26.26

the different images [9]. Despite the inherent stereo setup of this common two sensor deployment, in practical scenarios it is rarely exploited. A separate optical stereo setup is often favoured for the desired recovery of depth information [3]. However, the direct recovery of depth information from such a cross-spectral stereo setup[1] could facilitate stereo within the existing sensor foot-print (i.e. power, weight, cost, size, complexity) for applications such as obstacle avoidance [18], object detection [13] and tracking [21].

Prior work on the direct recovery of depth information from cross-spectral stereo images is limited [21, 22, 31]. Krotosky and Trivedi [21, 22] investigate cross-spectral stereo for pedestrian detection and tracking, using a window-based Mutual Information (MI) approach inspired by the original work of Egnal [10]. However, depth computation is only performed for isolated objects (i.e. pedestrians) via prior foreground extraction and subsequent localised stereo matching [21, 22]. Krotosky and Trivedi [22] additionally demonstrate the failure of dense depth computation using MI in the global energy minimisation framework of [14] caused by the lack of a global intensity transform between the images. The MI energy term cannot be effectively minimised globally as both good and bad matches produce similarly large values. More recently, Torabi and Bilodeau [31] describe a very similar window-based approach but replace MI by Local Self-Similarity (LSS) as a correspondence measure. LSS was originally proposed in [26] 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 [31]. Similar sparse scene feature matching techniques are proposed for the related problem of cross-spectral image registration [4] but these do not consider depth recovery.

Furthermore, the review of [16] identified a number of pre-processing filters and matching costs robust to the less challenging stereo problem of inter-image radiometric (illumination) differences. A number of authors [10, 12, 11, 19] propose using variants on an MI based approach for cross-spectral stereo matching based on results achieved on purely simulated cross-spectral data (i.e. where one image has undergone a radiometric transform to simulate an infrared/thermal image or similar). Here we illustrate both the limited applicability of these simulated data results [10, 12, 11, 19, 16], and the limited performance of approaches from prior cross-spectral work [21, 22, 31], in comparison to the use of dense gradient features with an appropriate optimisation approach.

2 Proposed Approach

Following the taxonomy of [25], dense stereo matching approaches can be split into four steps:- 1) pixel matching cost computation, 2) cost aggregation, 3) disparity (depth) optimisation and 4) disparity refinement and post-processing. Here we will concentrate on step 1-3 whilst assuming established post-processing approaches (step 4, [25]). The reader is directed to [25] for a general overview of dense computational stereo.

2.1 Cross-Spectral Rectification

In order to facilitate successful stereo matching the problem of cross-spectral stereo calibration has to be addressed. In prior work this sub-problem has either been avoided via the use of simulated imagery (with prior calibration in the visual spectrum) [12, 11, 19] or via scene feature driven registration [4]. Here we utilise the established calibration approach of [33] with a dual-material calibration target which is visible in both spectra (i.e. metal plate

[1] The terms multi-spectral, cross-spectral or multi-modal are often used interchangeably to generally refer to systems combining images in different spectral bands. Here we will consistently use the term cross-spectral for the combination of standard optical (visible) and thermal (infrared) images.

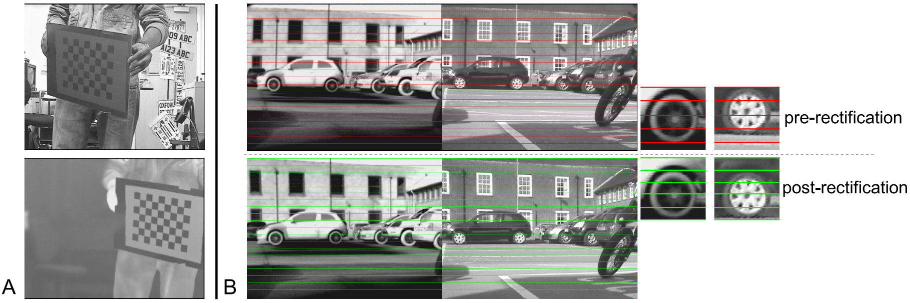

Figure 1: Cross-spectral calibration target (A) and rectified result (B)

with overlain reflective fabric "chessboard" pattern). This is simply heated using hot air or a high-power halogen lamp such that the varying latent heat properties of each material maximise the separation within the far infrared (thermal) spectral range required (see Figure 1A). Based on the approach of [33] the imagery is undistorted and rectified using the intrinsic and extrinsic camera parameters respectively (see Figure 1B). In general, robust cross-spectral calibration is identified as an area for future work while here we show that the use of [33] with a suitably engineered calibration target is viable.

2.2 Cross-Spectral Pixel Matching

While there is clearly no direct relation between pixel intensity values in this case, as are exploited by standard stereo algorithms, obvious similarities exist on a semantic level considering objects and object boundaries. Many corresponding object boundaries and edge fragments appear in both spectra, enabling a human observer to easily match corresponding objects in the images (e.g. Figure 1). From this observation we motivate our approach of using statistical local shape features based on image gradient orientations as a dense correspondence measure. This concept is commonplace in the feature descriptor approaches of SURF [2], SIFT [23] and alike. However, in cross-spectral images the orientation of image gradients do not correspond unambiguously because bright regions in the visible image can be dark in the thermal image and *vice-versa* (this is both sensor and ambient temperature dependent, Figure 1). As a result, we base our similarity on the unsigned gradient orientation, i.e. always mapping to the interval $(0,\pi)$, as proposed for object detection in [8]. This Histogram of Oriented Gradient (HOG) approach creates a descriptor optimised for "dense robust coding of spatial form" [8] robust to radiometric and illumination changes for object detection that can both be efficiently computed and readily compared using the L1 or L2 distance between descriptors.

Our Histograms of Oriented Gradient (HOG) features are a variant of the approach proposed by [8]. 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). 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 [8]. To maximise invariance we normalise the whole descriptor to L2 unit norm with the resulting HOG descriptor as a $n \times n \times H$ descriptor per pixel. A comparison, hence matching cost, between any two HOG descriptors is thus computed using the L1 distance. Dense HOG descriptors for every image pixel are computed efficiently by using integral histograms [24] allowing fast descriptor computation but preventing the use of spatial weighting (e.g. Gaussian) of gradient responses within any given descriptor in this case.

2.3 Disparity Optimisation

In addition to robustly computing the localised matching cost, the quality of the overall disparity image depends heavily on the disparity optimisation method used [25, 16]. Compared to standard stereo images, cross-spectral images can be expected to produce more ambiguous or false matches as well as weaker correct matches. Weaker correct matches can be caused by the difficulty of matching cost metrics to cope with naturally different appearance in the different spectra (e.g. see Figure 1B). It is thus important to compare how different optimisation techniques can compensate for these difficulties within a cross-spectral context. Overall a set of five such optimisation techniques are considered [25, 29, 15] and applied to the computed matching cost volume.

The simplest ("textbook") method which we investigate first is the Winner-Takes-All (WTA) method where the disparity producing the minimum matching cost is chosen at each pixel location [25]. The next, somewhat more advanced method, is a Dynamic Programming (DP) approach which enforces additional constraints along the image rows and is computationally efficient [25]. In addition, we test Scan-line Optimisation (SO), a common variation of dynamic programming which in contrast to a regular DP approach does not explicitly account for occlusions or enforce an ordering constraint [25]. Furthermore, we include Hirschmueller's seminal Semi-Global Matching (SGM) [15] which is both computationally efficient and provides improved global disparity smoothness constraints compared to DP. Finally, we evaluate the performance of global optimisation using Graph Cuts (GC) (expansion-move) to ascertain if improved results are achievable at additional computational cost [6, 5, 20, 29].

To enforce additional local smoothness constraints and reduce matching cost outliers we combine SO and DP optimisation with adaptive cost aggregation similar to [32] and apply a simple equally weighted (results in Section 3.1) or Gaussian weighted (results in Section 3.2) box filter to the matching costs prior to WTA, SGM and GC optimisation.

3 Results

For comparative evaluation within the context of prior work in this area [10, 19, 12, 11, 16, 21, 22, 31] we reproduce results from [10, 11, 16] on simulated cross-spectral imagery in addition to evaluating both a method based on recent LSS feature driven work [26, 31] and our own dense gradient feature proposal on the same (Figures 2 - 4). From this we down-select a subset of approaches [10, 16, 31] for comparison to the proposed approach on true cross-spectral imagery (Figures 5 - 10).

3.1 Simulated Cross-Spectral Stereo

In Figure 2 (left) we show the "parking meter" (upper) and "shrub" (lower) stereo pairs from the CMU VASC image database [1]. Furthermore we illustrate the optical stereo result achieved using Zero Mean Normalised Cross Correlation (ZNCC) [16] and in addition show a transformed version of the left stereo image of each pair (Figure 2, right) following the simulated transforms of [12, 11].

Using these simulated (right = *visible*, left = *transformed "infrared"*) stereo pairs we evaluate the proposed radiometric invariant approaches of *ZNCC*, Zero mean Sum of Absolute Differences (*ZSAD*), Rank based matching and Census based matching from [16] and in addition, further following [16], ZNCC variants with standard mean, Laplacian of Gaussian (*LoG*), background subtraction by *Bilateral* filtering and *Gradient Magnitude* response

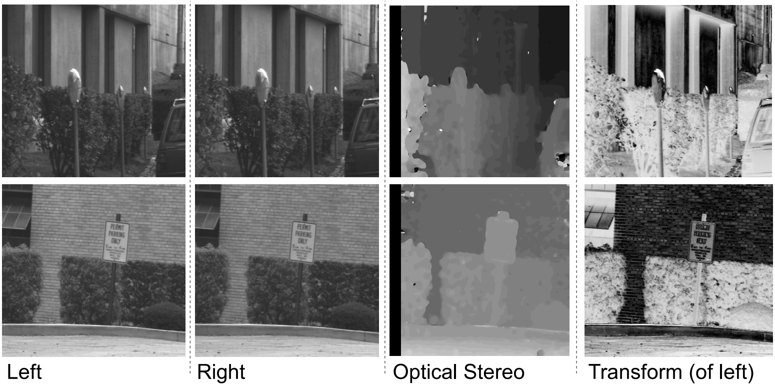

Figure 2: Example imagery for evaluation on simulated cross-spectral stereo

image pre-processing (Figure 3, [28]). All are computed with a simple Winner-Takes-All (WTA) disparity selection while other parameters are set with reference to the original works (*ZNCC/ZSAD/Rank/MI* $w = 11$; *Census* $w = 7$; *MI #bins* $= 16$, $\lambda = 0.4$; *LSS patch size = 5, region size = 35, log-polar grid = (4 radial, 12 angular)* [10, 11, 16, 31]) and HOG parameters ($H = 9$, $n = 3$, $b = 18$) specific to this task [8].

In Figure 3 we can see that, with the exception of ZNCC (gradient magnitude), the resulting depth images from these techniques are largely invalid and un-interpretable for any form of further use in scene understanding.

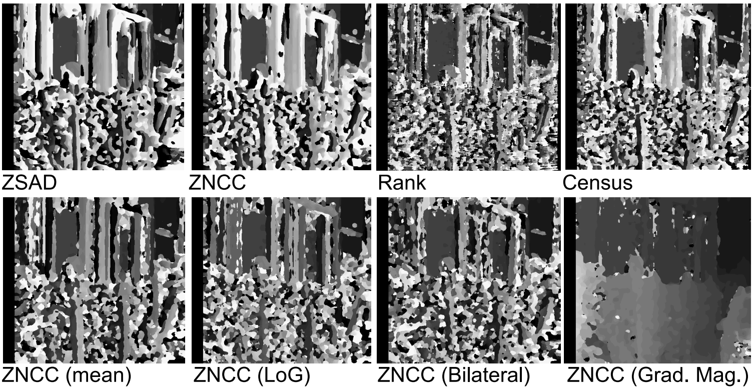

Figure 3: Basic cost matching approaches applied to simulated cross-spectral stereo

Furthermore, we evaluate variants of both the regular and the hierarchical (fixed window size) Mutual Information (MI) techniques of [10] and [12, 11] and a dense Local Self-Similarity (LSS) approach [31] (LSS descriptors used analogous to our HOG proposal) on both of these simulated cross-spectral stereo pairs together with our own HOG approach (Figure 4). In Figure 4 we see, for both "parking meter" (upper) and "shrub" (lower), a set of results comparable both to the performance of ZNCC (gradient magnitude) and to the reference optical stereo depth image (Figure 2, centre right). Dense gradient features are shown to at least match the performance of prior work based on this evaluation over simulated cross-spectral stereo imagery.

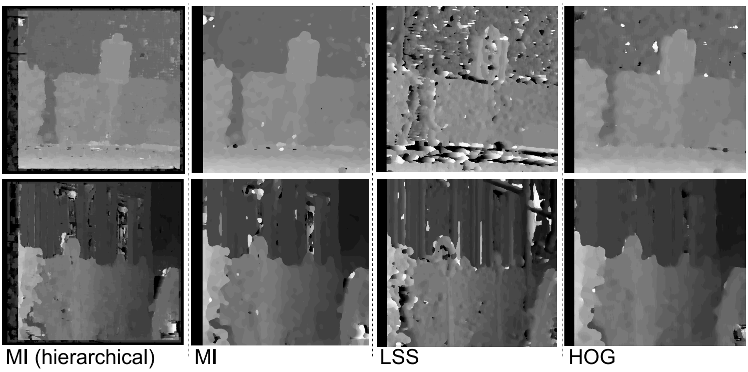

Figure 4: Advanced cost matching approaches applied to simulated cross-spectral stereo

3.2 True Cross-Spectral Stereo

We show the results of our proposed approach using a stereo rig consisting of an un-cooled far infrared camera (*Thermoteknix Miricle 307k*, spectral range: 8-12μm) and an optical vision camera (*Visionhitech VC57WD-24*, spectral range: ~400-700nm). Both cameras provide imagery at 640x480 resolution and are mounted on a mobile platform. All parameters are set as in Section 3.1 apart from *ZNCC/MI* $w = 21$ *and MI* $\lambda = 0.3$ (see [16, 11]). Results are illustrated over a range of scenes, in varying conditions and compared over variants of matching feature utilisation and optimisation (Sections 2.2 - 2.3).

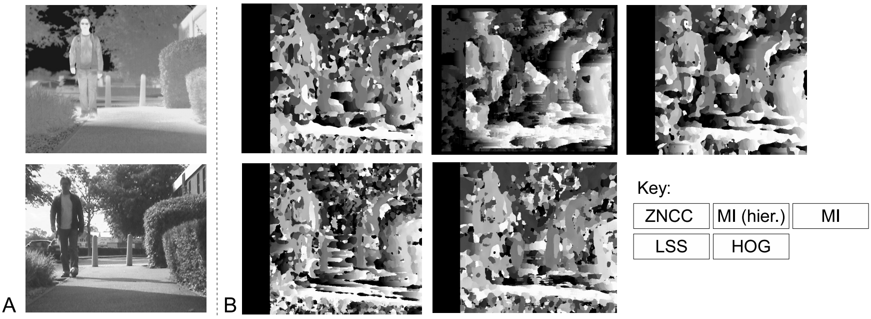

Figure 5: Cross-spectral stereo imagery (A) and calculated depth results (B)

Figure 5A shows a cross-spectral stereo scene *(stereo left image = top, stereo right image = bottom)* from which Figure 5B shows the depth image results obtained with each of ZNCC (gradient pre-processing) [16], MI [10], hierarchical MI [11], LSS [31] and our proposed approach with a simple WTA selection approach. The contrast between these results (Figure 5B) and those of the same techniques on the simulated imagery in Figure 4 is notable, suggesting true cross-spectral matching is significantly more challenging than the simulated case considered to date [19, 10, 12, 11].

Figure 6 shows the depth images obtained from the same cross-spectral stereo scene (i.e. Figure 5A) using HOG matching costs with each of the SO, DP, GC and SGM optimisation approaches. From the results on this scene (Figures 5 and 6) we can see that although all of the matching approaches perform poorly without optimisation (i.e. WTA, Figure 5B), it is possible to improve this performance under optimisation using strong smoothness constraints (Figure 6). This results in the recovery of varying levels of coherent scene depth suitable for scene understanding and reasoning. Pairwise we see the SO/DP techniques [25] and

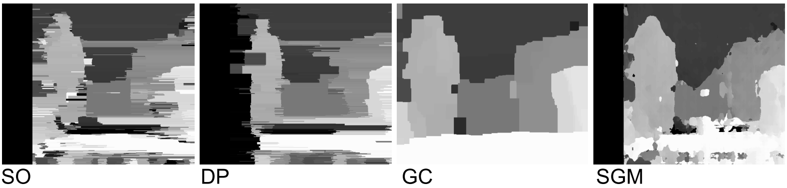

Figure 6: Cross-spectral stereo depth results with varying optimisation approaches

GC/SGM [29, 15] techniques producing somewhat similar results (Figure 6).

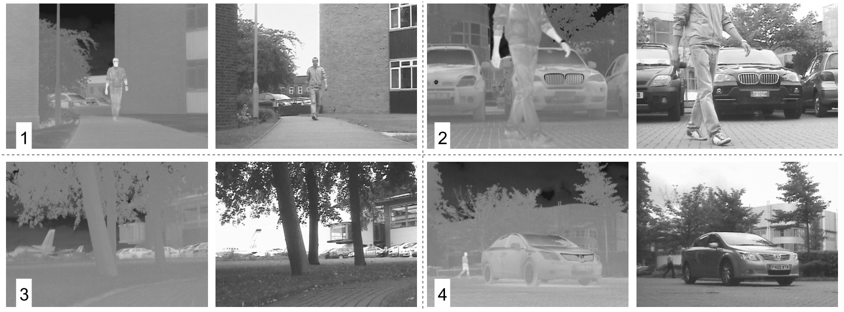

Figure 7: Cross-spectral evaluation scenes 1-4

From our initial results of Figures 5 / 6 we further compare the use of ZNCC (gradient) [16], MI [10], LSS [31] and the proposed dense HOG approach under two optimisation approaches (DP / SGM [25]). Whilst representative of the results of Figure 6, these are potentially achievable within real-time performance bounds [25, 14]. This comparison is performed over the four cross-spectral stereo scenes shown in Figure 7 with the results for scenes 1 and 2 shown in Figure 8 and those for scenes 3 and 4 shown in Figure 9. In the absence of explicit ground truth we base our evaluation on qualitative comparison of :- a) the cohesivity, connectedness and clarity of the resulting depth images and b) a comparison to that achieved over the same scene using optical stereo between two visible band cameras mounted on the same stereo rig (using Census based matching costs [16]). This reference is designed to provide grounding to the cross-spectral results in terms of what is possible using an established visible-band stereo technique under the same scene conditions. In the results shown the disparity outputs of SGM are additionally post-processed using left-right consistency checking and speckle removal [15].

In Figures 8 and 9 we can see that the use of dense HOG features (bottom, highlighted red) generally outperforms the other approaches, under both of the DP and SGM optimisation approaches, based on our evaluation criterion. They also offer results that are most similar in quality to those achievable under the same conditions using regular optical stereo (top, highlighted blue). In general, the performance of the dense HOG features under SGM optimisation can be seen to offer clearer and more cohesive depth image results than the same with DP optimisation. Furthermore, combined HOG+SGM stereo facilitate results that are most similar to those achievable with the regular optical stereo. This is most notable in scene example 2 (Figure 8 right) and scene example 4 (Figure 9 right) where notably large foreground objects are present. To a lesser extent this can also be seen in scene example 1 (Figure 8 left) and scene example 3 (Figure 9 left).

Overall, Figures 6 - 9 show that dense gradient features, based on our proposed HOG descriptor approach, combined with a strong optimisation approach facilitate viable cross-

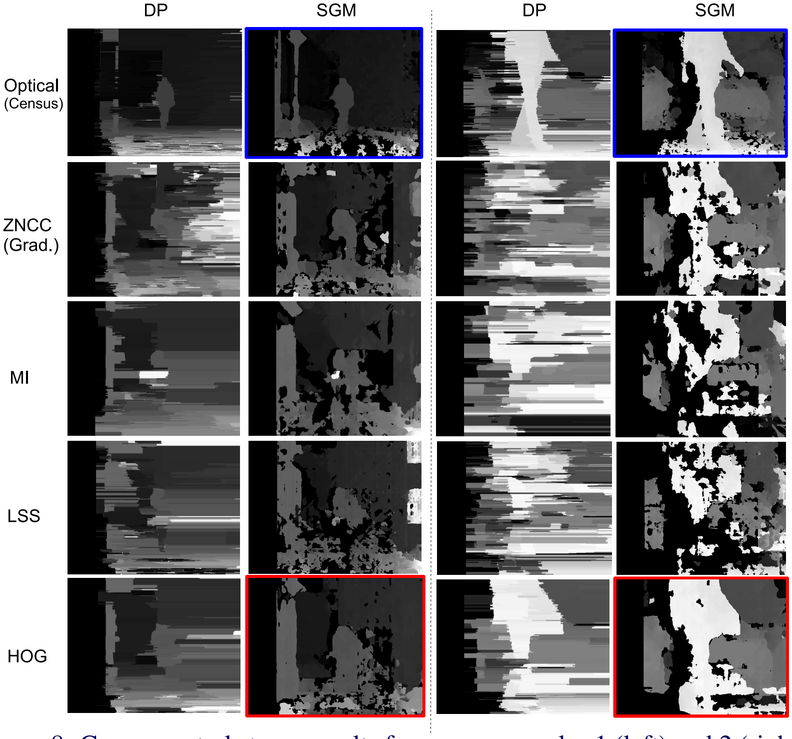

Figure 8: Cross-spectral stereo results for scene examples 1 (left) and 2 (right)

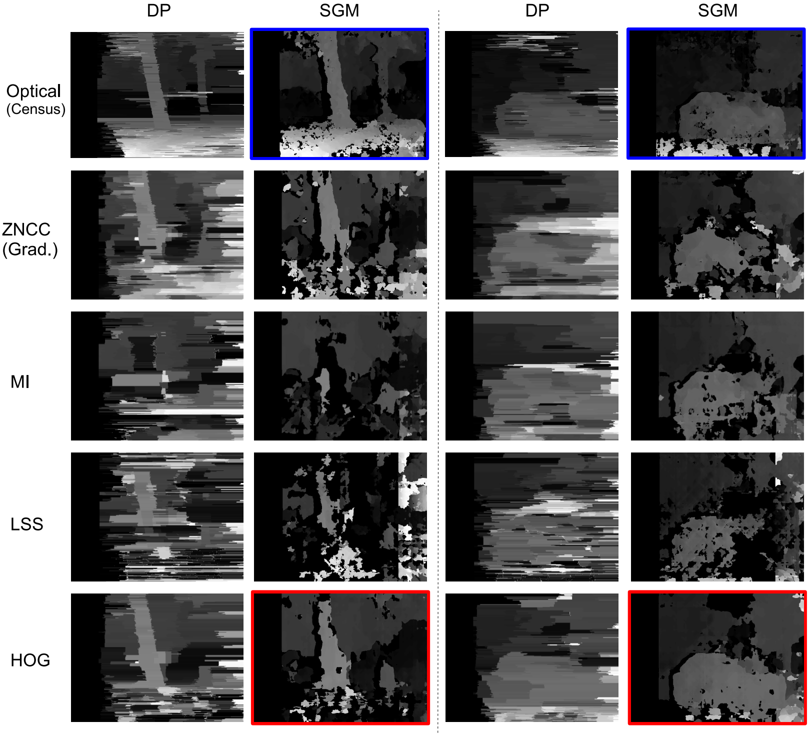

Figure 9: Cross-spectral stereo results for scene examples 3 (left) and 4 (right)

spectral stereo for the recovery of consistent (yet notably coarse) dense depth scene information. These depth image results could readily form the basis for further object localisation, obstacle avoidance and alike [13, 18] as part of a scene understanding approach which has hitherto been unavailable from a cross-spectral sensing arrangement. Whilst the authors fully accept that the rigour of this evaluation criteria falls short of the ground-truth comparison of the seminal Scharstein and Szeliski study within the field of conventional stereo [25], it is nevertheless apparent that the presented method outperforms previously proposed approaches [10, 11, 16, 31] for truly dense correspondence computation under the experimental conditions shown. Presently, the approaches implemented are not optimised (or parallelized) for run-time performance but indicative run-times for a single stereo pair (640×480) on a 2.4GHz Intel Core i5 CPU are:- MI (~360 s.), ZNCC (~120s.), LSS (~60s.) and HOG (~6s.) (for results shown in Figure 8 and 9).

3.3 Temporal Consistency

For completeness, we show temporally consistent depth is recovered over two illustrated sequences for objects visible in both camera views using combined HOG features and SGM (Figure 10 left and right, sub-sampled at $1Hz$ from $15fps$ video). This is achieved without explicit temporal consistency constraints.

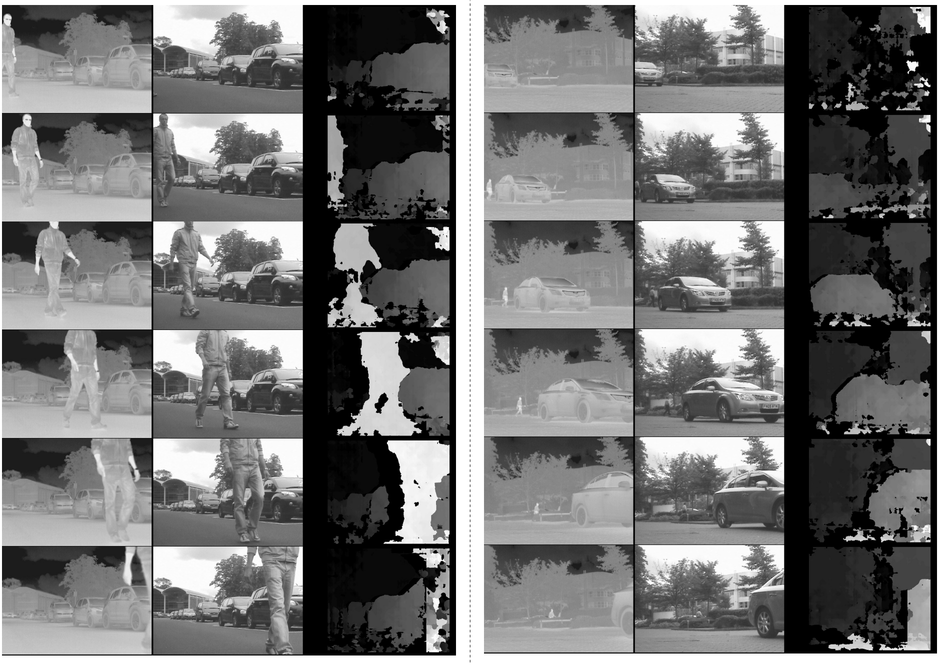

Figure 10: Cross-spectral stereo sequences with HOG+SGM depth recovery

4 Conclusions

Cross-spectral stereo matching can be achieved by using dense gradient features producing a scene depth image usable for further scene analysis and understanding. This extends prior work which is limited to simulated cross-spectral results [10, 19, 12, 11], image registration [4] or isolated object depth recovery [21, 22, 31]. By contrast, we show full scene depth recovery comparable in quality to standard optical stereo techniques under identical scene conditions. We illustrate that dense gradient feature approaches outperform methods based on prior work using Mutual Information (MI) [10, 12, 11] and Local Self-Similarity (LSS) features [31]. Furthermore, we show that prior results on radiometric image differences [16] or simulated imagery [10, 19, 12, 11] do not readily transfer to the true cross-spectral case. The prevalence of dense gradient approaches, notably dense Histograms of Oriented Gradient (HOG) features, over a range of disparity optimisation approaches is shown with improved results under strong optimisation criteria of Graph Cuts (GC) [6, 29] and Semi-Global Matching (SGM) [15]. Although the results remain somewhat coarse in comparison to contemporary work in optical stereo [16, 30], this work illustrates both :- a) the additional challenge of cross-spectral stereo in comparison to other stereo matching cases (e.g. radiometric differences [16]) and b) that results suitable for further scene analysis and understanding are achievable via a dense gradient feature approach. Future work will investigate both explicit evaluation against ground truth and consideration of further efficient dense gradient representations [30, 17] towards achieving real-time cross-spectral stereo.

References

[1] CMU Vision and Autonomous Systems Center (VASC) Image Database. URL http://vasc.ri.cmu.edu/idb/. accessed May 2012.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. ISSN 10773142. doi: 10.1016/j.cviu.2007.09.014.

[3] M. Bertozzi, A. Broggi, M. Felisa, S. Ghidoni, P. Grisleri, G. Vezzoni, C.H. Gómez, and M.D. Rose. Multi stereo-based pedestrian detection by daylight and far-infrared cameras. In R.I. Hammoud, editor, *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*, chapter 16, pages 371–401. Springer, 2009. ISBN 978-1-84800-276-0. doi: 10.1007/978-1-84800-277-7.

[4] C. Bodensteiner, W. Huebner, K. Juengling, J. Mueller, and M. Arens. Local multi-modal image matching based on self-similarity. In *Proceedings IEEE International Conference on Image Processing*, pages 937–940. IEEE, 2010. ISBN 978-1-4244-7992-4. doi: 10.1109/ICIP.2010.5651219.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.60.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001. ISSN 01628828. doi: 10.1109/34.969114.

[7] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del Solar. Human detection and identification by robots using thermal and visual information in domestic environments. *Journal of Intelligent & Robotic Systems*, 66(1):223–243, April 2012. ISSN 0921-0296. doi: 10.1007/s10846-011-9612-2.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.177.

[9] S. Denman, T. Lamb, C. Fookes, V. Chandran, and S. Sridharan. Multi-spectral fusion for surveillance systems. *Computers & Electrical Engineering*, 36(4):643–663, 2010. ISSN 00457906. doi: 10.1016/j.compeleceng.2008.11.011.

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

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

[12] 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. ISBN 0-7695-2223-8. doi: 10.1109/TDPVT.2004.1335420.

[13] A. Gaszczak, T.P. Breckon, and J.W. Han. Real-time people and vehicle detection from UAV imagery. In *Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, volume 7878, 2011. doi: 10.1117/12.876663.

[14] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 807–814. IEEE, 2005. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.56.

[15] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1166.

[16] 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. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.221.

[17] I. Katramados and T.P. Breckon. Real-time visual saliency by division of gaussians. In *Proc. International Conference on Image Processing*, pages 1741–1744. IEEE, September 2011. doi: 10.1109/ICIP.2011.6115785.

[18] I. Katramados, S. Crumpler, and T.P. Breckon. Real-time traversable surface detection by colour space fusion and temporal analysis. In *Proc. International Conference on Computer Vision Systems*, pages 265–274. Springer, 2009. doi: 10.1007/978-3-642-04667-4_27.

[19] J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *Proceedings Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1033–1040. IEEE, 2003. ISBN 0-7695-1950-4. doi: 10.1109/ICCV.2003.1238463.

[20] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.1262177.

[21] 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. ISSN 10773142. doi: 10.1016/j.cviu.2006.10.008.

[22] 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.

[23] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94.

[24] F. Porikli. Integral histogram: a fast way to extract histograms in Cartesian spaces. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 829–836. IEEE, 2005. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.188.

[25] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002. ISSN 0920-5691. doi: 10.1023/A:1014573219977.

[26] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. ISBN 1-4244-1179-3. doi: 10.1109/CVPR.2007.383198.

[27] D.A. Socolinsky. Design and deployment of visible-thermal biometric surveillance systems. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–2. IEEE, 2007. ISBN 1-4244-1179-3. doi: 10.1109/CVPR.2007.383531.

[28] C.J. Solomon and T.P. Breckon. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010. ISBN 0470844736. doi: 10.1002/9780470689776. ISBN-13: 978-0470844731.

[29] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (6):1068–1080, 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.70844.

[30] E. Tola, V. Lepetit, and P. Fua. DAISY: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.77.

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

[32] L. Wang, M. Liao, M. Gong, R. Yang, and D. Nister. High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In *Proceedings Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 798–805. IEEE, 2006. ISBN 0-7695-2825-2. doi: 10.1109/3DPVT.2006.75.

[33] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. ISSN 01628828. doi: 10.1109/34.888718.