On the Impact of Parallax Free Colour and Infrared Image Co-Registration to Fused Illumination Invariant Adaptive Background Modelling

Michael Loveday[1] and Toby P. Breckon[1,2]
[1]Department of Engineering | [2]Department of Computer Science
Durham University, UK

Abstract

Contrary to other visible-band (colour, RGB) and infrared-band (T) cross-modal work in the field, we present a practical approach to parallax-free RGB-T image formation using a combination of optical engineering (beam-splitter) and visual geometry. We use background to foreground object separation, a task inherently susceptible to multi-view parallax issues, to illustrate our approach. We evaluate the complementary nature of visible and far infrared (thermal, long-wave) information through three fusion schemes which physically combine visible-band (colour, RGB) and infrared-band (T) imagery into a co-registered, parallax free RGB-T image model. The performance of this combined RGB-T image model is assessed against standalone colour and thermal imagery for object detection within an adaptive background modelling framework. Illumination invariant background models, incorporating additional infrared information, increase the accuracy and precision of foreground object detection by over 10% on average when compared to standalone visible-band and over 5% for standalone infrared. Furthermore, the use of combined colour and infrared within adaptive background modelling provides superior results under conditions when either visible or infrared band performance is notably degraded. Evaluation is performed over a range of challenging conditions, over which the combined use of infrared and illumination invariant colour emerges as a more robust background modelling approach.

1. Introduction

The use of both colour and infrared imagery within many visual surveillance tasks is well established with numerous solutions spanning target detection, visual tracking and behaviour analytics [31, 20]. Central to almost all of these approaches is the use of background modelling [6, 26] for the initial segmentation of moving scene objects from the static scene background - the concept of *foreground* object detection from the modelling of a *scene background*. A wide and varied set of approaches exist for this task which has been extensively evaluated both in terms of colour [6, 26] and infrared (thermal) imagery [9, 32, 20, 21].

More broadly, the complementary nature of using both colour and (far, long-wave) infrared has seen them extensively utilised in a range of computer vision applications for several decades [23]. Prior work on dual colour-infrared (RGB-T) imagery has spanned object tracking [7, 40], pedestrian detection [19], face recognition [4] and applications within autonomous platform deployment [3]. Further work has also addressed the challenge of cross-spectral stereo between colour and infrared (thermal) stereo pairs [27, 1, 41] and presents extensions of popular feature matching approaches into this multi-modal space [2, 29, 16]. More recently we have seen a range of approaches using a dual colour and infrared camera setup to tackle sensing challenges within platform autonomy such as wide area search [3], 3D scene mapping [43] and Simultaneous Localisation and Mapping (SLAM) [24, 5].

Figure 1. Exemplar - a background model (left) and extracted foreground object (right) using the HST fusion model.

Parallel to this theme of using a broader spectral range for scene understanding tasks, we also find the literature littered with decades of work on numerous approaches for visible and infrared band image fusion for the end goal of single image presentation to a human viewer [44]. The reader is directed to the recent comprehensive reviews of [14, 22] for further insight.

Overall, this extensive body of prior work generally relies on a two camera side-by-side hardware setup [7, 3, 24, 43, 5, 14, 34] with an explicit cross-spectral image registration step to address image alignment between the colour and infrared (thermal) cameras. However, despite the use of even the most advanced cross-spectral registration approaches [30, 38], such a setup inherently introduces the problem of parallax within the scene [11] - *"the displacement or difference in the apparent position of an object viewed along two different lines of sight"* (one line of sight being from each of the cameras). Whilst this is exploited as the basis for cross-spectral stereo [10, 18, 27, 1, 16, 17, 41], it is often largely ignored within current detection, localisation and fusion RGB-T approaches [3, 43, 24, 34, 14] pre-

senting an ever present source of cross-spectral registration error that varies with object or feature depth in the scene [11]. Furthermore, a wide range of approaches that rely on the dual use of colour and infrared information, including for use in background modelling [34], either fail to present an explicit comparison of the performance gain from the use of additional infrared (thermal) band information or do so in the presence of the inherent parallax registration error [8, 39, 43, 24, 5].

By contrast, in this paper we leverage the work of [44, 25] to provide parallax-free colour (RGB) to infrared (thermal, T) registered imagery via the use of a gold dichroic beam-splitter to facilitate an orthogonal dual camera setup (Figure 2). This allows both image modalities to be captured to a common image projection plane via aligned optical axes - enabling parallax-free four-channel RGB-T formation. While use of such an approach has been established previously [44], it has received only limited attention within the broad scene understanding literature [42, 12, 15] with the work of [35] being most similar to our own. Within our formulation we further address issues of planar image alignment for final registration via established visual geometry (i.e. translational alignment of image centres, Section 2.2) and measure temporal synchronisation across varying modality camera hardware (Section 2.3). To illustrate our parallax-free approach within a visual sensing application context, here we explicitly present a comparison of the performance gain achieved with the use of infrared imagery combined with a number of illumination invariant colour representations, within the context of the adaptive background modelling [46] (Section 2.5). This is a task where inter-camera parallax would otherwise pose a significant issue - as the per-pixel foreground and background separation would vary from each viewpoint (i.e. modality). This is evaluated across a range of exemplar test scenarios designed to pose difficulties for each modality independently from which we see the combined use of invariant colour emerge as a more robust cross-modal background modelling approach (Section 3).

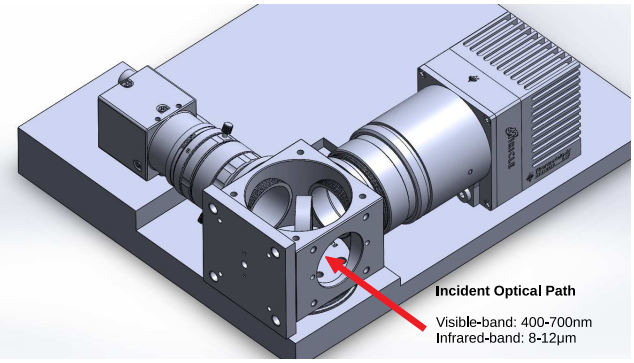

Figure 2. Dual-camera optical axis alignment via a beam-splitter (3D CAD representation incorporating components [28, 37, 13]).

2. Approach

To enable our parallax free RGB-T registration approach, we propose the use of optical image alignment for capture via aligned optical axes (Section 2.1) and field-of-view correction such that subsequent image registration is thus reduced to planar image alignment via a homography transform (Section 2.2). This is formulated within a measurable frame synchronisation error (Section 2.3), channel-wise infrared to colour fusion (Section 2.4) and an established adaptive background approach (Section 2.5).

2.1. Optical Alignment

To capture the optically aligned visible and infrared imagery, a *Thermoteknix Miricle 307k* un-cooled far infrared camera [37] (infrared-band: 8-12μm spectral range) and a *Point Grey Blackfly* [28] RGB colour camera (visible-band: 400-700nm spectral range) are used. The cameras are mounted perpendicularly using a gold dichroic beam-splitter [13] angled at 45° between them, as depicted in Figure 2, following the approach outlined in [25]. This beam-splitter reflects the 3-12μm spectrum towards the infrared camera while allowing transmission of the 400 - 700nm spectrum to the visible-band camera. The resulting imagery, a single channel infrared image with a corresponding three channel RGB colour image, is thus captured via aligned optical axis via the beam-splitter arrangement, hence avoiding the aforementioned parallax issue found in other work [39, 43, 24, 5].

2.2. Image Registration

The issue of cross-spectral calibration is tackled using the solution described by Pinggera et al. [27], by adapting the well-established method of Zhang [45] through the use of a calibration target visible in both modalities (a metal plate with a 'chessboard' pattern made from reflective material which is then heated before capturing images). The intrinsic camera parameters are determined via [45], through observing this planar calibration target at different orientations with refinement via Levenberg-Marquart optimisation.

Global field-of-view correction is performed, post image undistortion from intrinsic parameter correction, to account for differences in field of view between the infrared ($f_{infrared_h} = 28.1$, $f_{infrared_v} = 21.2$) and visible-band ($f_{visible_h} = 21.4$, $f_{visible_v} = 16.2$) cameras. To these ends, the infrared image is rescaled by the horisontal, ($\frac{f_{infrared_h}}{f_{visible_h}}$), and vertical field of view ratio, ($\frac{f_{infrared_v}}{f_{visible_v}}$), to match that of the visible band image [11]. After field-of-view correction, the images are aligned (registered) using a homography transform (Eqn. 1), calculated based on the same target based calibration method as before [45, 27]. The homography transform combines image rotation (θ), warping distortion (ω_1,ω_2), translation (t_x, t_y) and global scaling s to account for errors introduced from the beam-splitter not

being aligned at exactly $45°$.

$$H = \begin{bmatrix} \cos(\theta) & \sin(\theta) & t_x \\ -\sin(\theta) & \cos(\theta) & t_y \\ \omega_1 & \omega_1 & s \end{bmatrix} \quad (1)$$

In practice, the calibration determined the output rotation (θ), warping distortion (ω_1,ω_2) and scaling correction, s, are negligible returning an affine homography transform for image registration consisting of purely a residual translation (t_x, t_y) to align the images as would be expected within this optically aligned setup. Post image registration a common 640×480 resolution region is cropped from the registered imagery to form our combined RGB-T (four channel, visible + infrared) image.

2.3. Temporal Synchronisation

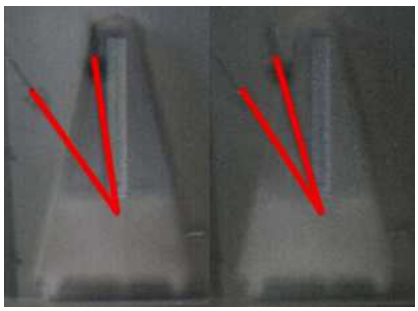

Figure 3. Overlain visible and infrared band for unsynchronised (left) and synchronised (right) image frames.

The remaining issue of temporal synchronisation between the frame buffers of the two diverse cameras within the sensing arrangement (Section 2.1), is tackled using a simple 555 timer circuit to externally trigger the cameras at $50Hz$ and hence provide a synchronised frame rate of $25fps$ (frames per second). Our cross-spectral hardware synchronisation is evaluated through the observation of a mechanical metronome (Figure 3), for which the simple harmonic motion the motion of the metronome is described by:-

$$\theta(t) = \theta_0 \sin\left(\frac{2\pi t}{T}\right) \quad (2)$$

where $\theta(t)$ is the angle with the vertical at time t, θ_0 the maximum angle, and T the period of the metronome. By rearranging Eqn. (2) the time between frames is given by:-

$$\Delta t = \frac{T}{2\pi}\left[\arcsin\left(\frac{\theta(t_1)}{\theta_0}\right) - \arcsin\left(\frac{\theta(t_2)}{\theta_0}\right)\right] \quad (3)$$

for the observed angle at times t_1 and t_2. The impact of the external synchronisation reduced the temporal synchronisation error from from $190ms$ to $77ms$ as can be observed between the unsynchronised (Figure 3, left) and synchonised infrared/colour image overlays (Figure 3, right - smaller angular difference in metronome position) . While this significantly reduced the temporal synchronisation error between the camera pair, optimally this delay should be less than the inter-frames interval ($40ms$ at $25fps$). As a result for fast object motion within the test imagery, the infrared frame visibly may lag behind the visible-band frame (e.g. Figure 7A) although $\sim 13fps$ is viable given the current synchronisation error. This remaining delay is partially attributable to the limited USB 2.0 bandwidth (single bus) connecting both cameras to the host computer and is treated as a known experimental error.

2.4. Colour and Infrared Fusion

In contrast to earlier fusion approaches in the field [14] we adopt a channel-wise fusion approach for our evaluation such that we retain the illumination invariant information from the colour RGB representation and insert an additional channel of infrared (T) information. Several visible-band colour-space models exist which represent colour information independently of illumination [33] which is itself well established in the literature as being detrimental to the adaptive background modelling task. Here we select the Hue, Saturation and Value (HSV) and Luminance, Chrominance (YCbCr) colour-space models such that illumination variation within the scene, corresponding to the V and Y channels respectively, can be removed and replaced with our infrared thermal scene information, denoted as T. In HSV colour-space the hue, H, and saturation, S, channels are retained. The illumination variation present within the V channel replaced with the corresponding infrared thermal image channel to form HST. Similarly, YCrCb describes the colour through two chroma components; red, Cr, and blue, Cb, with the luminance (illumination) information contained with the Y channel. We again replace this luminance (illumination) channel, Y, with the corresponding infrared thermal image channel to form $TCrCb$. Furthermore, we consider the use of $\{r, g\}$ chromaticity, as a normalised RGB derived colour measure known to be robust to changes in illumination calculated as $\{r, g, b\}$ as follows:

$$r = \frac{R}{R+B+G} \quad g = \frac{G}{R+B+G} \quad b = \frac{B}{R+B+G}$$

$$r + g + b = 1 \quad (4)$$

Comparatively, $\{r, g\}$ chromaticity normalises the RGB colour space and represents a colour through the proportion of the red, green and blue chroma present. Since these proportions are unit normalised, the blue value is commonly discarded for use in adaptive background modelling. The remaining $\{r, g\}$ chromaticity are combined with the corresponding infrared thermal image channel to form rgT.

All of these fused colour models, with retained illumination invariant colour and infrared information are illustrated

Figure 4. Exemplar RGB, infrared and varying visible-infrared fusion models (left to right).

in Figure 4 where we can see a visualisation of the resulting HST (Hue, Saturation, Thermal), $TCrCb$ (Thermal, Chroma-red, Chroma-blue), and rgT ($\{r, g\}$ chromaticity, Thermal) colour models in comparison to stand alone RGB and infrared (thermal, T).

Figure 5. Six reference scenarios used for evaluation.

2.5. Adaptive Background Modelling

Our reference adaptive background modelling approach is the mixture of Gaussian approach proposed by Zivkovic et al. [46] which is taken as highly representative of the relative performance other approaches in the field for the purposes of our evaluation [6, 26]. In this approach, a Gaussian Mixture Model (GMM) approach is used to model each pixel as proposed by [36]. The GMM model is adapted after each frame to allow for new stationary scene objects be incorporated into the background in addition to global/periodic scene changes based on a specified learning adaption rate. Whilst early GMM approaches used a fixed, number of Gaussian distributions within the model, Zivkovic [46] proposed an adaptive method which can automatically select the number of components needed per pixel and thus adapt more effectively to the observed scene. For our purposes a low learning rate (1×10^{-16}) is chosen to ensure stationary target objects are not incorporated into the background within the duration of the evaluation sequences. In addition, morphological opening and closing operations [33] are applied as a post-processing stage after initial foreground region separation via the GMM model in order to remove noise and reduce object holes which degrade overall object detection. The separation of foreground objects from the background using the approach is illustrated in Figure 1 where we can see the current background model from our HST fusion variant (left) and an isolated foreground pedestrian (right).

3. Evaluation

In order to evaluate the robustness of the varying RGB-T fusion models proposed, a number of surveillance sequences were gathered over a number of scenarios intended to challenge standalone visible-band or infrared object detection. These challenges include low lighting, shadows, change in illumination, small objects, scene clutter, similar background/foreground temperature, infrared reflection and thermal gradients within the scene.

3.1. Experimental Conditions

A total of eight dual RGB-T video sequences were captured in six different scene locations as depicted in Figure 5. These eight dual RGB-T video test sequences can be outlined as follows:

- ***Yard 1:*** set in a works yard with low lighting and a short subject distance to test detection under poor illumination. Pedestrian enters the scene, stops and makes a phone call and then exits towards the camera.
- ***Yard 2:*** set in a works yard but from an elevated position. Illumination is low and a puddle causes a reflection of the subject as it passes. Pedestrian enters and crosses the scene, stops before exiting the same way they entered.
- ***Alley:*** set at a road between two buildings where trees cause a number of shadows to be cast across the scene. Three sequences take place in this scene:-
  - ***Subject at a distance:*** pedestrian crosses the scene at a large distance from the camera to test small object detection.
  - ***Single object:*** pedestrian enters the scene and strolls away from the camera, stops and lingers before turning and exiting from via the same direction as entry. To tests detection of a single target in the presence of significant shadows.

– ***Multiple objects:*** multiple pedestrians enter the scene from different directions, two targets interact as they pass. To test multiple object detection and inter-occlusion.

- ***Car Park:*** set at a row of cars in a car park, with a shadow from a building causing a large change in illumination and temperature between the front and back of the scene. Pedestrian enters scene and crosses from shadow into sunlight and then stops in front of a 'hot' car in the sun before crossing back into shadow and leaving the scene.

- ***Trees:*** set at a road in front of a row of trees to test multiple object detection with oscillating background characteristics. Multiple pedestrians crossing the scene from multiple distances from the camera in addition to traversal of a vehicle target.

- ***Road:*** set at a road in bright illumination with trees casting varying shadows across the scene. Prolonged sunlight has warmed the background causing a similar background temperature to foreground objects within the scene. Pedestrian enters the scene from the rear (afar) and walks towards the camera and exits.

3.2. Statistical Results

Each of the outlined RGB-T fusion models, $\{HST, HCrCb, rgT\}$, are quantitatively evaluated by comparing the number of foreground objects detected in each image frame of each test video sequence. This is obtained via connected components analysis of the foreground image mask obtained from our adaptive background model, against manually annotated ground truth (i.e. actual foreground objects present). On this basis each frame is classified as either a true positive, tp, true negative, tn, false positive, fp, or false negative, fn, frame occurrence. A true positive frame is defined as one where all the foreground objects are correctly detected and true negative frame as one containing no foreground objects where no foreground objects are detected. Conversely, a false positive frame is defined where more foreground objects are detected than the ground truth records and similarly a false negative frame is one where one or more foreground objects present in the ground truth are not detected via the adaptive background model. An example of each of a tp, tn, fp, and fn frame occurrence is depicted in Figure 6.

Each RGB-T test video sequence is then evaluated in terms of per-frame statistical accuracy, precision, and recall, Eqn. (5) to Eqn. (7):-

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{5}$$

$$Precision = \frac{tp}{tp + fp} \tag{6}$$

$$Recall = \frac{tp}{tp + fn} \tag{7}$$

Within this statistical analysis, $Accuracy$ corresponds to the proportion of correct results among all of the image frames evaluated. $Precision$ describes the proportion of detected objects which were indeed correct, and $Recall$ represents the proportion of objects present that were correctly detected. The statistical results for each scene are presented in Table 1 for each of our outlined RGB-T fusion models, $\{HST, HCrCb, rgT\}$, in addition to each of stand-alone RGB colour and infrared (T) within the same adaptive background model formulation.

From Table 1 we can conclude that the primary source of error of all of the fusion schemes came from false positive foreground object detection as opposed to false negative detection - characterised by lower $Precision$ and higher $Recall$ with reference to Eqn. 6 & 7. A contributory factor to this is the detection of remaining background noise regions, which are not completely removed by morphological post-processing, as erroneous foreground objects (see the failure case in Figure 7A). Although the proposed RGB-T fusion models offer superior performance to regular visible-band colour (RGB) in all scenarios and infrared (T) in all but a single scenario, no single fusion model is conclusively superior across all the test scenarios. All of the proposed RGB-T fusion models outperform both standalone visible-band (RGB) and infrared (T) with respect to both accuracy and precision of foreground object detection, demonstrating the complementary nature of visible-band and infrared information within this context. Whilst on average the $TCrCb$ model performed marginally better across all scenarios, each fusion model exhibited strengths which are subjective to the scene.

3.3. Discussion

As expected visible-band detection performs poorly in comparison to infrared in the low illumination conditions of the *Yard 1* test sequence due to false detections (Table 1 - $Precision$). This is attributable to increased noise in the visible-band image which the morphological post-processing is unable to remove. The infrared provides the best performance, however the drop in effectiveness of the fusion models can be partially attributed to poor synchronisation between the cameras (Figure 7A). Similar results are observed over the *Yard 2* test sequence where low illumination causes detection errors and a reflection from a puddle consistently causes false detections to occur (Figure 7B). While being more robust to low lighting, rgT also detected the reflection demonstrating a stronger dependence for rgT on the visible-band information as opposed to the infrared. Both infrared and $TCrCb$ perform well with a small drop in performance due to an object oscillating in the background, however the addition of visible-band information notably

Figure 6. Exemplar foreground object detection for tp, tn, fp, and fn occurrences.

decreases $TCrCb$ susceptibility to this. HST is the worse at rejecting the oscillatory motion, however through its resilience to the reflection still outperforms both the visible-band and rgT.

The hybrid and infrared colour models are more adept at detecting small objects in the distance within the *Alley* test sequences. Here the visible-band model is often unable to detect a distant moving object, which is represented by particularly high false negatives (Table 1 - $Recall$). The more apparent difference in temperature between the foreground objects and the background enables a more effective detection for the infrared model, which combined with the colour information from the visible-band model facilitates superior detection for $TCrCb$ and rgT although not for HST. Contrastingly, when operating with a shorter object distance the large difference in contrast between the object and background from the sunlight improves the performance of the visible-band model. However, as the target moves into the sunlight, shadows are detected as foreground objects and degrade the performance slightly. Conversely, the infrared performance drops dramatically, producing fragmented objects and therefore is unable to recognise single objects instead reporting multiple detections per object over 50% of the time (e.g. Figure 7C). Parts of the target clothing are poorly detected by infrared and cause separate detections for the head and torso, as is visible in Figure 7C. Similarly to the *Yard 2* scene, rgT again shows correlation with the performance of the visible-band model, and through fusion with the infrared information outperforms all other models. In addition, the HST and $TCrCb$ shows a strong correlation with the infrared model since both also suffer from object fragmentation.

The transition between shadow and bright sunlight within the *Car Park sequence* causes both the visible-band and infrared models problems for different reasons (Table 1). In the case of the visible-band model, the target moves out from shadow where we should expect correct detection rates to increase. Although the bright sunlight at the rear of the scene back-lights the target and increases detection in the low section of the scene, when the target moves into the bright sunlight it becomes overexposed and is thus difficult to separate from the background (Figure 7D). Comparatively, while in the shadow the temperature between the target and the background are distinctly different allowing for accurate identification of the foreground object. However, the bright sunlight causes the ambient temperature of the background to increase, and therefore the target and background have similar thermal intensities and as a result reduce the effectiveness of background modelling (Figure 7E). By contrast the fusion models, using a combination of visible-band and infrared information, are able to compensate for the overexposure as well as the object and background temperature similarity (Table 1).

All of the approaches are suitably able to adapt to the oscillations within the *Trees* sequence (Table 1). However the introduction of large occlusions as objects move behind trees proves problematic for the visible-band colour model, this being worse for objects further from sensor as only small portions of the object are detected which are often mistakenly removed as noise (Figure 7F). The infrared, and fusion models, are clearly able to distinguish between the object and the background and as a result are more capable at correctly detecting objects even when partially occluded. On the other hand, the visible-band information is

Accuracy					
Scenario	*RGB*	*T*	*HST*	*TCrCb*	*rgT*
Yard 1	0.75	**0.99**	***0.98***	0.97	0.89
Yard 2	0.30	0.84	0.74	**0.90**	0.48
Alley 1	0.81	0.96	0.85	0.97	**0.98**
Alley 2	0.83	0.47	0.47	0.51	**0.90**
Alley 3	0.59	0.44	0.40	0.44	**0.63**
Car Park	0.70	0.71	**0.87**	0.80	0.80
Trees	0.67	0.72	**0.79**	0.74	0.78
Road	0.26	0.29	0.69	**0.75**	0.52
Mean	0.61	0.68	***0.72***	***0.76***	***0.75***

Precision					
Scenario	*RGB*	*T*	*HST*	*TCrCb*	*rgT*
Yard 1	0.18	**0.96**	0.91	0.88	0.60
Yard 2	0.26	0.84	0.73	**0.90**	0.45
Alley 1	0.95	0.96	0.84	0.95	**0.98**
Alley 2	0.79	0.35	0.35	0.40	**0.87**
Alley 3	0.43	0.22	0.18	0.23	**0.50**
Car Park	0.49	0.50	**0.78**	0.67	0.68
Trees	0.70	0.82	0.83	0.77	**0.85**
Road	0.26	0.24	0.66	**0.73**	0.52
Mean	0.51	0.61	***0.66***	***0.69***	***0.68***

Recall					
Scenario	*RGB*	*T*	*HST*	*TCrCb*	*rgT*
Yard 1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Yard 2	0.88	0.99	**1.00**	0.99	0.95
Alley 1	0.73	0.97	0.89	**0.99**	**0.99**
Alley 2	**1.00**	0.98	**1.00**	0.98	**1.00**
Alley 3	**0.97**	0.59	0.55	0.74	**0.92**
Car Park	0.95	**1.00**	**0.98**	0.97	0.95
Trees	0.89	0.84	**0.92**	0.91	0.89
Road	**1.00**	0.97	**1.00**	**1.00**	**1.00**
Mean	0.93	0.92	*0.92*	***0.95***	***0.96***

Table 1. Accuracy, precision, recall: foreground object detection

able to more accurately detect the vehicle as opposed to the infrared which is unable to discern it from the background. Once again the fusion schemes are able to exploit the benefits of each modality and suppress the drawbacks to outperform both. $TCrCb$ produces the worst detection rate of all fusion models (Table 1), again attributable to poor object detection due to target clothing (Figure 7G).

Within the *Road* test sequence, the similar background and object temperature cause the infrared sensor to fail to differentiate between foreground and background. This results in a large number of false detections since the more easily detectable aspects of the object, such as bare skin, are reported as separate objects (Figure 7H). The visible-band results are also very poor, stemming from a combination of shadows cast across the scene and high exposure areas. The high exposure causes bright aspects of the object to be missed as most colour information is lost (Figure 7I). Interestingly, the colour fusion models are capable of producing much higher detection rates than the poorly performing visible-band and infrared. While the shadows are still detected, rgT, HST and $TCrCb$ are able to dramatically improve the detection when compared to visible or infrared bands alone. These models are capable of using the colour information to improve the detection where infrared is unable to separate objects of similar foreground and background temperatures, while also reducing the effect of the overexposed visible-band with the thermal information (Table 1).

Overall, from Table 1, we can see that the primary source of error is from false positive detections as opposed to false negative detections (lower $Precision$, higher $Recall$). This is attributable to the general failures in foreground noise removal and object fragmentation illustrated in Figure 7. The performance of foreground object detection is invariant to the number of such objects present within the scene however increased inter-object occlusion impacted overall performance.

While the combination of the visible-band and infrared modalities improve the performance of foreground object detection across all scenes the drawbacks of each could not be entirely suppressed in all cases. In general, each fusion model shows a strong correlation in performance with either the visible or infrared band results. rgT is more correlated with the performance of the visible-band whilst HST and $TCrCb$ are more correlated with the infrared. This results in rgT offering superior performance when the visible-band performs well although it similarly suffers in the presence of visible-band related challenges such as shadows and reflections. Similarly, HST and $TCrCb$ correlation with infrared results in superior performance when the infrared performs well but is less effective when foreground objects have similar temperature to the background. On average $TCrCb$ performs best across all the scenarios presented although this performance remains situationally variant.

4. Conclusions

Overall, we extend the work of [44, 25] to evaluate a practical approach to parallax-free RGB-T image formation using a combination of optical engineering (beam-splitter) and visual geometry. Furthermore, we present an approach to measure cross-spectral temporal camera synchronisation. We frame our approach within the context of adaptive background modelling, as a representative visual sensing task where inter-camera parallax would otherwise be a significant issue, and report the relative performance of the combined use of infrared and illumination invariant colour components against regular single modality sensing. For future work, we will look into the potential use of this parallax-free RGB-T approach in the other cross-modal applications in the field [39, 43, 24, 5].

Figure 7. Exemplar foreground object detection - failure cases (A - I).

References

[1] F. Barrera. Multimodal stereo from thermal infrared and visible spectrum. *Electronic Letters on Computer Vision and Image Analysis*, 13(2), 2014.

[2] C. Bodensteiner, W. Huebner, K. Jüngling, J. Müller, and M. Arens. Local multi-modal image matching based on self-similarity. In *International Conference on Image Processing*, pages 937–940. IEEE, 2010.

[3] T. Breckon, A. Gaszczak, J. Han, M. Eichner, and S. Barnes. Multi-modal target detection for autonomous wide area search and surveillance. In *Proc. SPIE Emerging Technologies in Security and Defence: Unmanned Sensor Systems*, volume 8899, pages 1–19, September 2013.

[4] H. Chang, A. Koschan, M. Abidi, S. G. Kong, and C.-H. Won. Multispectral visible and infrared imaging for face recognition. In *Proc. Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE, 2008.

[5] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan. RGB-T SLAM: A flexible SLAM framework by combining appearance and thermal information. In *Proc International Conference on Robotics and Automation*, pages 5682–5687. IEEE, may 2017.

[6] S. K. Choudhury, P. K. Sa, S. Bakshi, and B. Majhi. An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios. *IEEE Access*, 4:6133–6150, 2016.

[7] S. Colantonio, M. Benvenuti, M. Di Bono, G. Pieri, and O. Salvetti. Object tracking in a stereo and infrared vision system. *Infrared physics and technology*, 49(3):266–271, 2007.

[8] 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, 2012.

[9] J. W. Davis and V. Sharma. Background-subtraction in thermal imagery using contour saliency. *Int. Journal of Computer Vision*, 71(2):161–181, 2007.

[10] 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.

[11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proc. Conf. on Computer Vision and Pattern Recognition*, 2015.

[13] ISP Optics. Gold dichroic beamsplitter, 2016.

[14] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He. A survey of infrared and visual image fusion methods. *Infrared Physics & Technology*, 85:478–501, 2017.

[15] N. Kim, Y. Choi, S. Hwang, K. Park, J. S. Yoon, and I. S. Kweon. Geometrical calibration of multispectral calibration. In *Proc. Int. Conf. Ubiquitous Robots and Ambient Intelligence*, pages 384–385. IEEE, 2015.

[16] S. Kim, D. Min, B. Ham, S. Ryu, M. Do, and K. Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 2103–2112, 2015.

[17] S. Kim, D. Min, S. Lin, and K. Sohn. Deep self-correlation descriptor for dense cross-modal correspondence. In *Proc. European Conf. on Computer Vision*, pages 679–695. Springer, 2016.

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

[19] S. J. Krotosky and M. M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):619–629, 2007.

[20] M. Kundegorski and T. Breckon. A photogrammetric approach for real-time 3D localization and tracking of pedestrians in monocular infrared imagery. In *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, volume 9253, page (to appear). SPIE, sep 2014.

[21] M. E. Kundegorski, S. Akcay, G. de La Garanderie, and T. P. Breckon. Real-time Classification of Vehicle Types within Infra-red Imagery. In *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, volume 9995, pages 1–16. SPIE, sep 2016.

[22] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin. Pixel-level image fusion: A survey of the state of the art. *Information Fusion*, 33:100–112, 2017.

[23] S. S. Lin. Review: Extending visible band computer vision techniques to infrared band images. Technical report, University of Pennsylvania, 2001.

[24] M. Magnabosco and T. Breckon. Cross-Spectral Visual Simultaneous Localization And Mapping (SLAM) with Sensor Handover. *Robotics and Autonomous Systems*, 63(2):195–208, feb 2013.

[25] N. J. Morris, S. Avidan, W. Matusik, and H. Pfister. Statistics of infrared images. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.

[26] S. Panahi, S. Sheikhi, S. Hadadan, and N. Gheissari. Evaluation of background subtraction methods. In *Computing: Techniques and Applications*, pages 357–364. IEEE, 2008.

[27] P. Pinggera, T. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *Proc. British Machine Vision Conference*, pages 526.1–526.12. BMVA, September 2012.

[28] Point Grey. Blackfly 0.5 MP Color USB3 Vision (Sony ICX693), 2016.

[29] S. Saleem and R. Sablatnig. A robust sift descriptor for multispectral images. *IEEE Signal Processing Letters*, 4(21):400–403, 2014.

[30] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. In *Proc. European Conference on Computer Vision*, pages 309–324. Springer, 2014.

[31] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual Tracking: An Experimental Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, jul 2014.

[32] A. Sobral and A. Vacavant. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122:4–21, may 2014.

[33] C. Solomon and T. Breckon. *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010. ISBN-13: 978-0470844731.

[34] P. St-Charles, G. Bilodeau, and R. Bergevin. Mutual foreground segmentation with multispectral stereo pairs. In *Proc. Int. Conference on Computer Vision Workshops*, pages 375–384, 2017.

[35] L. St-Laurent, X. Maldague, and D. Prevost. Combination of colour and thermal sensors for enhanced object detection. In *Proc. Int. Conf. Information Fusion*, pages 1–8. IEEE, 2007.

[36] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 246–252, 1999.

[37] Thermoteknix Systems Ltd. Miricle Thermal Imaging Modules, 2013.

[38] 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*, pages 61–67, 2011.

[39] A. Torabi and G.-A. Bilodeau. A LSS-based Registration Of Stereo Thermal-Visible Videos Of Multiple People Using Belief Propagation. *Computer Vision and Image Understanding*, 117(12), jul 2013.

[40] A. Torabi, G. Massé, and G.-A. Bilodeau. An iterative integrated framework for thermalâvisible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221, feb 2012.

[41] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu. CATS: A Color and Thermal Stereo Benchmark. In *Proc. Conf. on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017.

[42] J. Van Baar, P. Beardsley, M. Pollefeys, and M. Gross. Sensor fusion for depth estimation, including tof and thermal sensors. In *Proc. Int. Conf. 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 472–478. IEEE, 2012.

[43] S. Vidas, P. Moghadam, and M. Bosse. 3D thermal mapping of building interiors using an RGB-D and thermal camera. In *Proc. Int. Conf. on Robotics and Automation*, pages 2311–2318. IEEE, may 2013.

[44] A. Waxman, M. Aguilar, D. Fay, D. Ireland, and J. Racamato. Solid-state color night vision: fusion of low-light visible and thermal infrared imagery. *Lincoln Laboratory Journal*, 11(1):41–60, 1998.

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

[46] Z. Zivkovic and F. Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006.