A Photogrammetric Approach for Real-time 3D Localization and Tracking of Pedestrians in Monocular Infrared Imagery

Mikolaj E. Kundegorski, Toby P. Breckon
School of Engineering and Computing Sciences, Durham University, UK

ABSTRACT

Target tracking within conventional video imagery poses a significant challenge that is increasingly being addressed via complex algorithmic solutions. The complexity of this problem can be fundamentally attributed to the ambiguity associated with actual 3D scene position of a given tracked object in relation to its observed position in 2D image space. We propose an approach that challenges the current trend in complex tracking solutions by addressing this fundamental ambiguity head-on. In contrast to prior work in the field, we leverage the key advantages of thermal-band infrared (IR) imagery for the pedestrian localization to show that robust localization and foreground target separation, afforded via such imagery, facilities accurate 3D position estimation to within the error bounds of conventional Global Position System (GPS) positioning. This work investigates the accuracy of classical photogrammetry, within the context of current target detection and classification techniques, as a means of recovering the true 3D position of pedestrian targets within the scene. Based on photogrammetric estimation of target position, we then illustrate the efficiency of regular Kalman filter based tracking operating on actual 3D pedestrian scene trajectories. We present both a statistical and experimental analysis of the associated errors of this approach in addition to real-time 3D pedestrian tracking using monocular infrared (IR) imagery from a thermal-band camera.

Keywords: thermal target tracking, temporal filtering, intelligent target reporting, thermal imaging, pedestrian detection, people detection, sensor networks, temporal fusion, passive target positioning, 3D pedestrian localization

1. INTRODUCTION

Contemporary approaches to visual target tracking are commonly addressed by increasingly complex algorithmic solutions [1]. This complexity is directly attributable to the fundamental ambiguity associated with actual 3D scene position of a given scene object in relation to its observed position in 2D image space. The complexity arises from the dual ambiguity in both 2D object position in the image (i.e. which image pixels constitute the target) and 3D object position within the scene in relation to the camera sensor (i.e. are we observing a small object, near field or a large object, far field ?).

Within the context of pedestrian tracking, we demonstrate that reasonable performance can practically be achieved through the combined use of infrared imagery (thermal-band, spectral range: 8-12μm) and the application of real-time photogrammetry. We leverage the key advantage of such thermal-band infrared (IR) imagery for the pedestrian target localization within the image (e.g. Figure 8). This facilitates both robust detection of human signatures within the scene [2–4] and robust localization of their scene bounds in pixel-space. We can use the principles of photogrammetry to recover 3D object position within the scene based on a known camera projection model and an assumption on the real-world size of our targets along a single dimension. The use of photogrammetry within this context is largely over-looked within the current literature with only limited utilization within a target localization and tracking context. Despite this, recent statistical studies strongly support the validity of the principle that underpins the assumption upon which this approach rests - that variance in human height is in fact quite small [5, 6].

Here we experimentally investigate the accuracy of classical photogrammetry, within the context of current target detection and classification techniques [2–4], as a means of recovering the true 3D position of pedestrian targets within the scene. We present a real-time approach for the detection, classification and localization of pedestrian targets via thermal-band (infrared) sensing suitable for use in a deployed network of autonomous sensor nodes.

Despite extensive work in ground-based sensor networks [7–10], the use of photogrammetry within this context has received only limited attention [11]. The visible-band work of [11] uses a similar approach within a Bayesian 3D tracking framework but does not explicitly address issues of accuracy or its use within a detection filtering framework [2]. In addition, some general scene understanding approaches have also used this

Corresponding author: toby.breckon@durham.ac.uk

principle to determine relative object dimensions and positions within the scene [12, 13] although alternative approaches such as active sensing [14], structure from motion [15] and monocular depth recovery [16, 17] have become increasingly popular within this task of late.

Prior work explicitly dealing with thermal-band (IR) imagery within an automated surveillance context is presently largely focused upon pedestrian detection [2, 4, 18–20] and tracking [21, 22]. More recently extended studies have investigated the fundamentals of both background scene modeling [23] and feature point descriptors [24] that commonly form the basis of many such techniques [2, 20]. Related work has begun to address the challenges of cross-modal stereo [25, 26] as a future aspect of the cross-spectral sensing solutions commonly deployed within autonomous sensing solutions [2, 4]. The work presented in this paper is a direct extension of [2] to consider photogrammetric target localization from a single modality ([2] presents multi-modal target detection) incorporating a lightweight tracking solution akin to that of [3]. Furthermore it provides both statistical and experimental validation of the photogrammetric assumptions on average human height that underpin the prior 3D tracking work of [11] and the general scene understanding approaches of [12, 13].

In contrast to prior work in the field, we investigate the accuracy now afforded to photogrammetric pedestrian target localization based on a) the key advantage of reduced pixel-space localization ambiguity within thermal-band infrared imagery and b) recent statistical results that report narrow standard deviations within large-scale surveys of human height variation. These combined aspects of increased confidence in pedestrian target dimensions identified within pixel-space [2] and increased confidence of marginal error in any assumption of corresponding real-world pedestrian height (i.e. average human height, [5, 6]) facilitate robust localization accuracy to within the commonly regarded *"gold-standard"* of consumer-level Global Position System (GPS) positioning (typically $\pm 5m$ under ideal conditions [27]). This is achieved using solely passive sensing from a monocular infrared imaging camera, with no *a priori* environment calibration. Within the context of prior work [2], this is presented in-conjunction with real-time detection, tracking and reporting of pedestrian surveillance targets with are localized to global scene position relative to a GPS-enabled camera sensor. This provides true 3D pedestrian tracking in a given environment akin to [11]. Results are presented over a range of evaluation scenarios using an evaluation methodology based upon both quantification of the error present within the photogrammetric localization approach proposed and a range of successful pedestrian tracking scenarios.

2. PEDESTRIAN TARGET LOCALIZATION

We perform localization, and subsequent tracking in real-world 3D space (*"scene space"*), based on the initial detection (Section 2.1), photogrammetric based localization (Section 2.2) and Kalman filter driven tracking over these recovered 3D position estimations (Section 2.3).

2.1 Pedestrian Detection

Our approach is illustrated against the backdrop of a classical two stage automated visual surveillance approach. First we detect initial candidate regions within the scene (Section 2.1.1), thus facilitating efficient feature extraction over isolated scene regions, to which an identified target type is assigned via secondary object classification (Section 2.1.2) [2].

2.1.1 Candidate Region Detection

In order to facilitate overall real-time performance, initial candidate region detection identifies isolated regions of interest within the scene. This allows subsequent feature extraction and classification to be performed over isolated region(s) enabling real-time processing. Additionally, this facilitates efficient object localization within the scene. By leveraging the stationary position of our sensor, this is achieved using a combination of two adaptive background modeling approaches [28, 29] working in parallel to produce a single robust foreground model over varying environmental conditions and notably within varying ambient thermal/infrared illumination conditions within complex, cluttered environments.

Within the first model, a Mixture of Gaussian (MoG) based adaptive background model, each image pixel is modeled as a set of Gaussian distributions, commonly termed as a Gaussian mixture model, that capture both noise related and periodic (i.e. vibration, movement) changes in pixel intensity at each and every location within the image over time [28, 30]. This background model is adaptively updated with each frame received and each pixel is probabilistically evaluated as being either part of the scene foreground or background following this methodology. The second model comprises the use of Bayesian classification in a closed feedback loop with Kalman filtered predictions of foreground component position [29]. Within this model, each pixel is similarly probabilistically classified as either foreground or

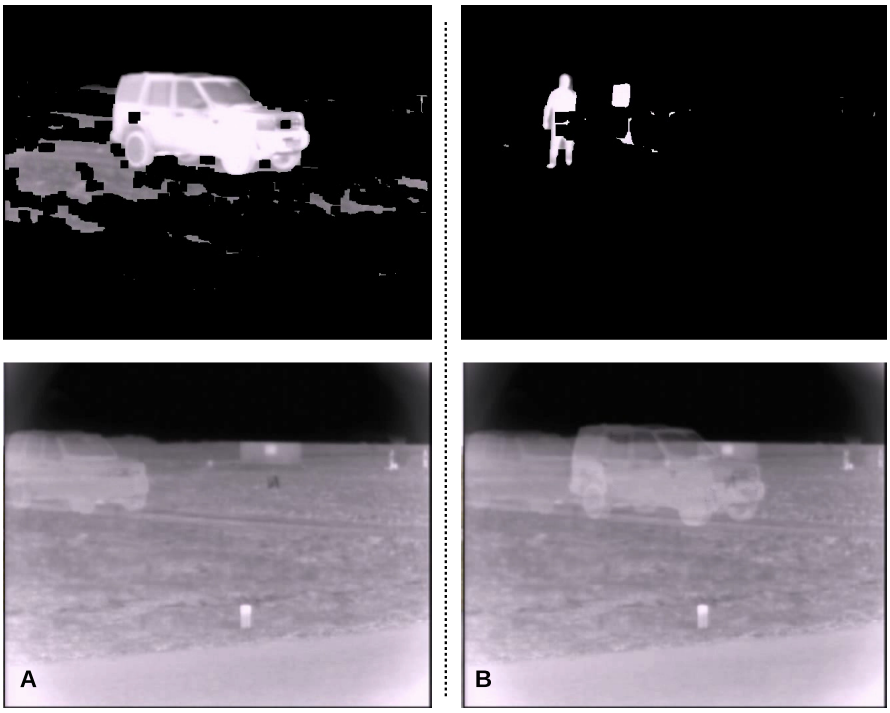

Figure 1. Candidate regions (upper) detected against current MoG background model (lower) [2]

background but this is further reinforced via Kalman predictions for the positions of foreground objects (i.e. connected component foreground regions [31]) present in the previous time-step. This object-aware model significantly aids in the recovery of fast moving foreground objects under varying illumination conditions such as the thermal gradients inherent within infrared imagery.

Overall this combined approach provides a slowly-adapting background model in the traditional sense [28], that can be robust to rapid illumination gradients, whilst similarly providing foreground consistency to fast moving scene objects [29]. The binary output of each foreground, based on a probabilistic classification threshold, is combined conjunctively to provide robust detection of both static and active scene objects.

Within our overall framework this facilitates the automated identification of foreground regions within the incoming video imagery such that a) new objects can be isolated from the scene background for efficient feature generation and classification and b) objects that enter and become static within the scene (e.g. parked car) are adaptively learnt as part of the background model. This concept is illustrated in Figure 1 where we see isolated scene regions relating corresponding to the vehicle and pedestrians entering the scene (Figure 1 A/B upper). These scene regions (Figure 1 A/B upper) represent pixel values that do not fit the current adaptive background model and are thus identified as foreground pixels within the image. This set of foreground pixels (Figure 1 A/B upper) is post-processed using morphological dilation and connected components analysis [31] facilitating the rejection of small noisy candidate regions prior to classification. It is notable that some small noise regions remain and similarly noise remains within the foreground object boundaries (Figure 1 A/B upper).

In Figure 1 A/B (lower) we see a visualization of the current background model in each instance based on a weighted average of the current Gaussian mixture model at each scene pixel. Figure 1B also shows how the representation of the stationary vehicle that was present in the scene for some time (Figure 1A) is incorporated into the background model and subsequently removed as it departs the scene following the continuous updates to the adaptive background model itself. Traces of the vehicles prior position in the scene background are clearly visible in Figure 1B (lower) relating to its previous position in Figure 1A. Minor traces of the vehicles previous stationary position (maintained shortly after entering the scene) are also visible in Figure 1A (lower) but have subsequently been removed by updating in Figure 1B which is taken from later in the

same sequence (once the vehicle has departed).

Overall, this background modeling approach facilitates the efficient identification of candidate regions for further feature extraction and classification (Section 2.1.2). As a by-product it readily facilitates the reporting of new, arriving and transiting/moving scene objects as scene events without continual re-reporting of stationary scene objects that were not originally present within the scene. The use of such adaptive background modeling techniques is commonplace in the automated visual surveillance and tracking literature [30].

2.1.2 Target Classification

Target classification follows a machine learning driven approach of off-line classifier training and on-line target classification using the trained classifier. Specifically we follow the bag of visual words methodology [32–34]. Following this approach abstract multi-dimensional visual features are extracted over the set of training examples. In general a wide range of features are available and widely utilised for this task [35, 36] with variation in computational performance, complexity and invariance properties. Here we utilise the Speeded Up Robust Features (SURF) approach [37] that are both viable for real-time performance on full motion video and known to be suitable for multi-modal use as our base features [20, 24].

Following the bag of visual words (or codebook) methodology, we perform multi-dimensional feature clustering over all of the example training imagery (for all object classes) to produce a set of general feature clusters that characterise the overall feature space. This provides a fixed dimension set of cluster references for all target types and sub-types that we are to classify. Commonly this set of feature clusters is referred to as a codebook or vocabulary as it is subsequently used to encode the features detected on specific object instances (positive and negative) as fixed length vectors for input to both the off-line training and on-line classification phase of later machine learning classification. Here we perform clustering using the common-place k-means clustering algorithm in 128-dimensional space (i.e. SURF feature descriptor length of 128 [37]) into 1000 clusters. A given object instance is encoded as a fixed length vector based on the membership of the features detected within the object to a given feature cluster based on nearest neighbour (hard) cluster assignment. Essentially the original variable number of SURF features detected over each training image or candidate region is encoded as a fixed length histogram representing the membership of these features to each of these clusters. This fixed length distribution of features forms a feature vector that is then used to differentiate between positive and negative instances of a given class based on a trained classifier. The feature vector forms the input to a two-class Support Vector Machine (SVM) classifier, $pedestrian = \{yes, no\}$, that is trained using a RBF kernel, via grid-based kernel parameter optimization, within a cross-validation based training regime [38].

2.2 Photogrammetric Position Estimation

Based on automated detection (Section 2.1.2), target position is initially known within *"sensor space"* (i.e. pixel position within the image). Consequently, target position is estimated based on the principles of photogrammetry together with knowledge of the perspective transform under which targets are imaged and an assumption on the physical (real-world) dimension of a target in one plane.

All targets are imaged under a standard perspective projection [31] as follows:

$$x = f\frac{X}{Z},\ y = f\frac{Y}{Z} \tag{1}$$

where real-world object position, (X, Y, Z), in 3D scene co-ordinate space is imaged at image pixel position, (x, y), in pixel co-ordinate space for a given camera focal length, f. We assume both positions are the centroid of the object with (x, y) being the centre of the bounding box, of the image sub-region, for a target (object) detected in the scene (Section 2.1.1, e.g. Figure 2).

With knowledge of the camera focal length, f, the original object (target) position, (X, Y, Z), can be recovered based on (assumed) knowledge of either object width, $\triangle X$, or object height, $\triangle Y$ (i.e. the difference in minimum and maximum positions in each of these dimensions for the object). From the bounds of the detected targets (Section 2.1.2) we can readily recover the corresponding object width, $\triangle x$, and object height, $\triangle y$, in the image. Based on this knowledge, rearranging and substituting into Eqn. 1 we can recover the depth (distance to target, Z) of the object position as follows:

$$Z = f'\frac{\triangle Y}{\triangle y} \tag{2}$$

Knowing Z via Eqn. 2, we can now substitute back into Eqn. 1 and with knowledge of the object centroid in the image, (x, y), we can recover both X and Y resulting in full recovery of real-world target position,

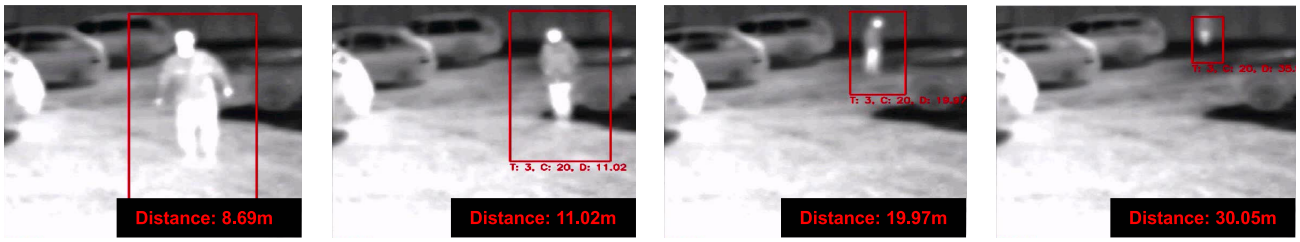

Figure 2. Photogrammetry facilitates the approximate recovery of a camera to target distance for an example target (person) without any need for additional (active) range sensing [2]

(X, Y, Z), relative to the camera. In Eqn. 2, f' represents focal length, f, translated from standard units, mm, to focal length measured in pixels:-

$$f' = \frac{width_{image} \cdot f}{width_{sensor}} \tag{3}$$

where $width_{image}$ represents the width of the image (pixels), $width_{sensor}$ represents the camera CCD sensor width (mm).

Crucially, if we now assume a fixed width, $\triangle X$, or height, $\triangle Y$, for our object we can recover complete 3D scene position relative to the camera. For pedestrian detection we can assume average adult human height based on available medical statistics [5, 6]. Despite commonly held beliefs, notable large-scale studies have shown variance on human height within the adult population to be low ("*in populations of European descent, the average height is ∼178 cm for males and ∼165 cm for females, with a standard deviation of ∼7 cm*" [6]). The meta-study of [6] considers a total populace of 63,000 individuals within its analysis from several studies over which average variation present translates into a very narrow ∼3.9-4.2% height difference across adults (for each gender) and an average difference of 8% between the sexes. It is estimated that approximately 80% of this variation is due to as few as 50 genetic factors [39]. This means that height variation can be expected to be very small within the general populace. Hypothetically, if all of the genetic factors causing most of this variation were known and summed together, it is estimated that the (extreme case) height variation between the upper-most 5% (tallest) and lowest 5% percentile (shortest) within the population would be ∼26cm [6] in that case (i.e. the height of a human head).

Within our work, this variance is directly proportional to the recovered object depth estimate, Z (and similarly to (X, Y)) via Eqn. 2. Despite the crudeness of this assumption, empirically it has been shown to work well within the context of target localization in earlier work [4, 11]. This is supported by assessing the error effect of the variation identified by [6] within this context (Figure 3). Figure 3 (left) shows the position error in the Z position estimate that would result from a 7cm standard deviation in height, $\triangle Y$, within Eqn. 2 for either male, female and the combined adult population. It can be seen that this translates to a maximal Z position error of ∼2.5m at a 60m range (linearly scaling to ∼5m at a 120m range, equivalent to established GPS error tolerances under ideal conditions [27]). For extreme cases of height variation within adults (i.e. upper/lower 5% outliers within the population under hypothetical conditions) we see a maximal error of ∼9m at a 60m range and an error which is still within established GPS error tolerances for ranges <33m (Figure 3, right). These extreme cases of variation are highly unlikely to occur with great regularity, based upon widespread statistical surveys [5, 6].

Although this analysis considers the adult height only, it is widely accepted that a 14-year old male is at ∼90% of his full adult height with this percentage being even greater for females of this age [40]. Based on the mean height of a 14-year old male (164cm) for an individual going on to reach average male height (178cm), adding this additional potential source of height variation (+7.9%) will only result in a position error of approaching ∼5m (established GPS error tolerances [27]) at a 43m range (Figure 3 (right)). By contrast, for females (14-year old, 156cm [40], +2.4% height variation), this translates as a position error <4m for targets within 60m range (Figure 3 (right)) as females are typically already at ∼95+% of full adult height by this age. Similar analysis can be performed for varying age and gender combinations based on [40].

Extending our analysis to consider the distribution of height variation across the population, based on the full adult height distribution presented within the extensive study of [40], we can consider the Z position error introduced due to height variation in the populace between the 0.4 - 99.6 percentile (Figure 4). Based on a mean of 178 cm for males and 165 cm for females, we plot

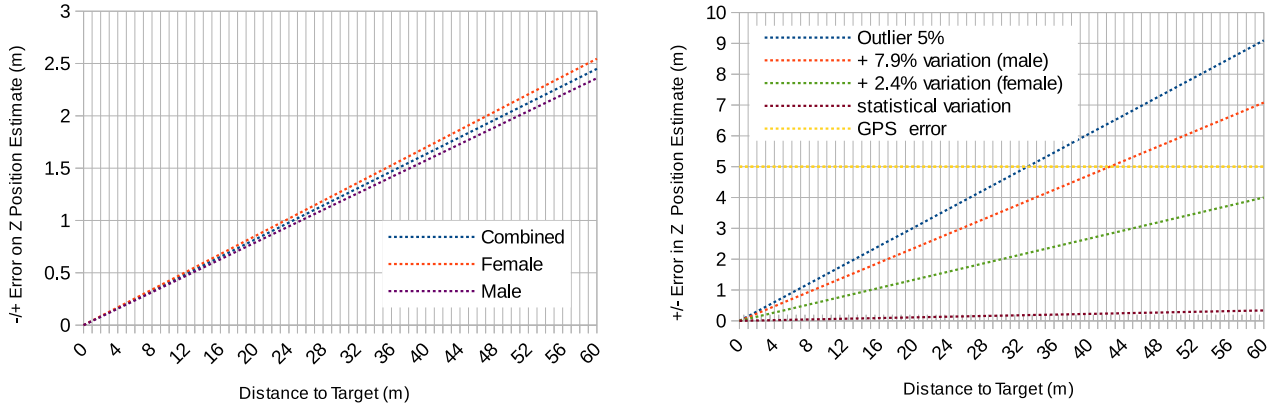

Figure 3. Position error in Z (distance to target) attributable to variation in height for standard deviation of 7cm (left) and for cases of additional height variation due to genetics and age differences (right)

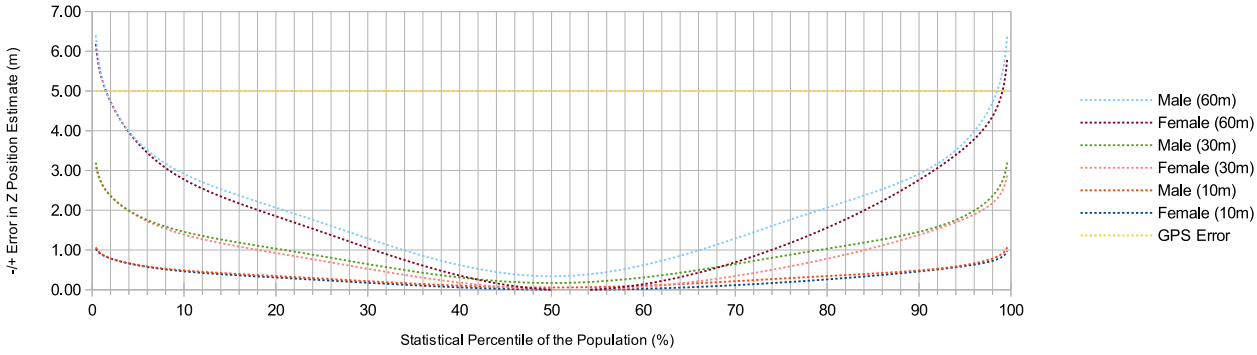

Figure 4. Position error in Z (distance to target) attributable to variation in height across the 0.4 →99.6 percentile of the population [40].

the Z position error introduced at each of {10m, 30m, 60m} target distances due to this distribution height variation in Figure 4. We can conclude that height variation at 10m or 30m distances introduces a Z position error within GPS error tolerances for the entire populace (male and female) (Figure 4). Furthermore the Z position error, attributable to height variation, at 60m distance is within GPS error tolerances for approximately the ∼2-98% percentile of the population (based on height distribution). Essentially, the photogrammetric approach we propose can be statistically shown to introduce an error that exceeds standard GPS position error tolerances for less than 4% of individuals (Figure 4, [40]).

Variation within the various statistical sources used for height information in this analysis [5, 6, 40] and in those used within prior work [4, 13] result in at most a variation in reference height of ±1cm. This statistical variation is itself shown to have a negligible effect on position error relative to other error sources and established GPS error tolerances [27] (see Figure 3 (right)).

Figure 2 illustrates the application of this approach to the position estimation, showing distance to target only, with an example pedestrian target that is detected using the approach outlined in [4]. It is similarly shown in Figure 5 (right) using the approach used here (Section 2.1.1). Within the earlier work [4] we can additionally see a secondary source of error present - the estimate of object height, $\triangle y$, in the image. This is estimated based on the bounding box of the detected individual (Figure 2) using either the full height dimension, a fixed percentage thereof (Figure 2, [4]) or a secondary process of extended limb localization within the bounding box [3]. Within [4] this introduces a notable secondary source of position estimation error (Eqn. 2) as the bounding box may be a poor approximation to the actual height of the individual within the image (Figure 2, right →middle). In Figure 5, we explicitly compare this earlier cascaded Haar classifier driven approach of [4] with the approach outlined here in Section 2.1.1 [2] over the same image sequence. In general, our chosen approach (Figure 5, right) produces a tighter bound on the target with greater consistency over the duration of the sequence. This minimizes, although does not eliminate, this source of error on $\triangle y$ in the image. Within our evaluation (Section 3) we experimentally examine the impact of this remaining $\triangle y$ error upon the overall Z position estimation of targets and show empirically that its effect is small (Figures 6 / 7).

Overall, we can effectively show that strong statistical support for a photogrammetric approach capable of delivering Z position estimates to statistical tolerances within current GPS accuracy in the majority of instances [5, 6, 40]. This is further supported by a target localization approach within the image, that minimizes error introduced by poor height bounds on targets within the image (Section 2.1.1, [2]). Furthermore it crucially offers a passive, as opposed to active sensing, based position estimation for detected targets. Based on a sensor position that is itself known *a priori* (from on-board GPS or mapping) and target position relative to the sensor recovered using this approach, (X, Y, Z) (Eqn. 2), is readily transformed into global position coordinates for onward target reporting within the observed environment. This is illustrated for a range of pedestrian targets and environments within Figures 8 - 10.

2.3 3D Tracking

Unlike conventional tracking approaches that track 2D position, (x, y), within the image itself [1, 41], our photogrammetric recovery of target position within the scene, (X, Y, Z) (Section 2.2) facilitates 3D tracking within scene space. This can be accomplished as tracking *"within the plane"* based on horizontal target position within the scene, X, and distance to target, Z, or full 3D scene space tracking including target elevation (vertical position), Y. Whilst any detected foreground object (Section 2.1.1) can be tracked based on 2D position, we require confirmed classification as a pedestrian target (Section 2.2).

For each candidate region identified as a new foreground object (Section 2.1.1), we initially created a new 2D track-let based on localized frame to frame connectivity derived from sparse optic flow [42, 43]. If one of the frame samples for this object is subsequently classified as pedestrian (via the approach outlined in Section 2.1.2), this target transitions from a 2D tracked instance within image space to a 3D tracked pedestrian within scene space. The tracked position, based on photogrammetric position recovery (Section 2.2) can then be propagated, over earlier instances of the same object similarly transitioning the motion history of this instance from 2D image position to 3D scene position. If an identified foreground object is not classified as being a pedestrian its tracking remains within 2D image space until either its spatio-temporal filtered classification (Section 2.3.1, Eqn(s). 4 & 5) returns a pedestrian classification or it leaves the scene. Tracking within 3D scene space is performed using Kalman filter based tracking [44] on either a state vector comprising position and velocity *"within the plane"*, $\vec{s} = (X, Z, vX, vZ)^T$, or within $\mathbb{R}^3$ scene

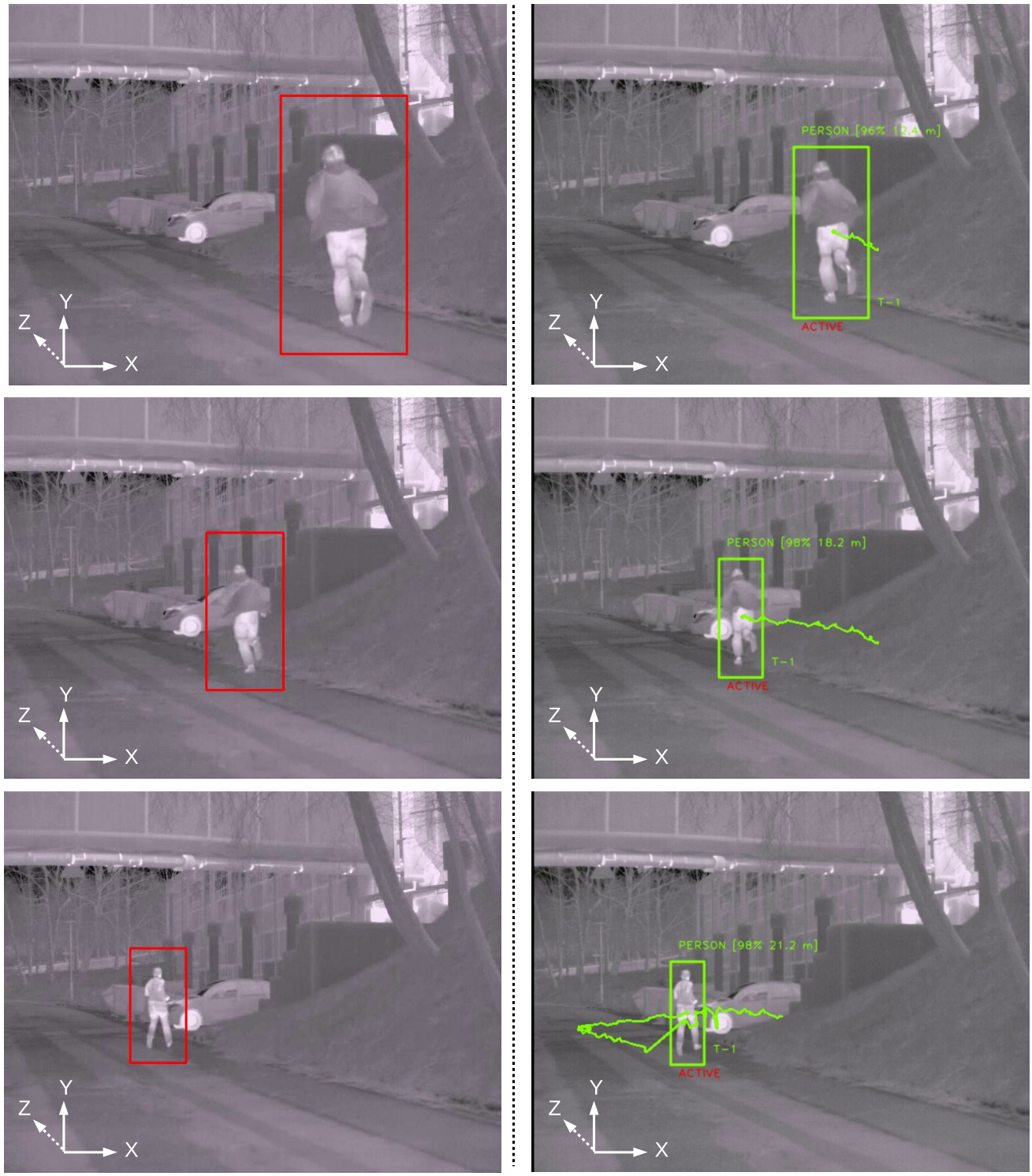

Figure 5. A comparison of the localization of target bounds using the prior approach of [4] and that used here (Section 2.1.1, [2]).

space, $\vec{s} = (X, Y, Z, vX, vY, vZ)^T$. Scene and measurement noise within the Kalman formulation are estimated empirically.

2.3.1 Spatio-temporal Target Filtering

Adapting the temporal filtering approach of [2], 3D tracking allows us to perform target filtering both spatially (as 3D tracked objects) and temporally (along the duration of the track itself). In this way we can similarly make use of weak classifiers following the discussion outline in [2] placing significantly less reliance on a single well-trained classifier that performs well under all conditions.

For a classifier to be truly weak its output must be correlated, albeit poorly, against true target detection. Despite this inherent weakness, we can make use of it by performing a bagging or boosting ensemble classification approach over a given time (sampling) window [2]. False positives or false negatives will be random and not temporally or spatially clustered over any significant sampling period or spatial area. By contrast, true positive (i.e. correct) detections will be temporally and spatially clustered. This key observation allows us to proceed with a less than optimal (weaker) classifier approach and rely on the strength of an ensemble approach to facilitate the desired reduction in false positive reporting.

For a classifier detecting the presence of target class c based on a given feature distribution x, represented by a binary classification function $k_c(x) \in \{0, 1\}$, this results in a classification result integrated and normalized over a temporal window of w at time t as follows:

$$A_{k_c} = \frac{\sum_{t-w}^{t} k_c(x_t)}{w} \tag{4}$$

Here we spatially constrain this temporal window, on a per scene object basis (from tracking, Section 2.3), to be along the last w connected instances within a given object track in 3D scene space. Essentially we take the average classification, $k_c(x)$, over the last w tracked instances of a given object. This results in a real-valued integrated classification in the range $\{0 \rightarrow 1\}$ which if treated akin to a probability can be considered to give a likelihood of detection (normalized to a percentage with the results of Section 3). Applying a threshold to this parameter, τ_k, facilitates the translation of this to a detection report, $pedestrain = \{yes|no\}$, for onward transmission within the sensor network. In essence, as Eqn. 4 gives equal weighting to each time step within the temporal window, this can be directly translated to state - *if target is detected greater than τ_k% of the time within w sequential image samples we then report it as a confirmed target detection.* An extension would be to weight each time step decaying by time or by spatial proximity to the sensor (c.f. Section 2.2).

Our formulation (Eqn. 4) can be further expanded as follows to consider a non-binary classification function, $f_c(x)$, as follows:

$$A_{f_c} = \frac{\sum_{t-w}^{t} f_c(x_t)}{\beta w} \tag{5}$$

where β is the maximal value of $f_c(x)$ and thus A_{f_c}can be treated analogous to A_{k_c}. In practice our non-binary classification function may itself return a probability directly from an underlying Bayesian classifier or perhaps the number of positive votes within a decision forest classifier (where $\beta = \#trees\, in\, forest$) or the distance of the feature distribution instance from the decision boundary in Support Vector Machine (SVM) classification (where $\beta = max(f_c(i)) \forall\, i,\, i \in \{training\, examples\}$) [2].

Overall the use of spatio-temporal filtering constrains any target detections via classification (Section 2.1.2) both in terms of spatial consistency and temporal consistency. Prior work has experimentally shown that temporal filtering significantly reduces spurious false-positives [2] whilst here we introduce additional spatial constraint by simply piggy-backing off prior target tracking in 3D scene space.

3. EVALUATION

Our results are presented using both quantitative measures of pedestrian localization accuracy (Figures 6 / 7) and qualitative assessment of 3D localization and tracking performance over a range of exemplar scenarios (Figures 8 - 10). All evaluation imagery is captured using an un-cooled infrared camera (*Thermoteknix Miricle 307k*, spectral range: 8-12μm).

Figure 6 presents the quantitative results of target localization (i.e. estimated Z position, Section 2.2) averaged over multiple sample experiments plotted against ground truth for the ranges 10→30m (Fig. 6A) and 10→60m (Fig. 6B). Both sets of experiments (Fig. 6A / Fig. 6B) were carried out under different experimental conditions and in different locations. Error bars are plotted for the standard deviation in the Z position estimates obtained (y-axes, Figure 6) and for the expected error in range due to human height variation at this range (x-axes, Figure 6, derived from Figure 3 / Eqn. 2) in addition to $\pm 5m$ GPS error margins [27]. From these results it can be observed that the estimated

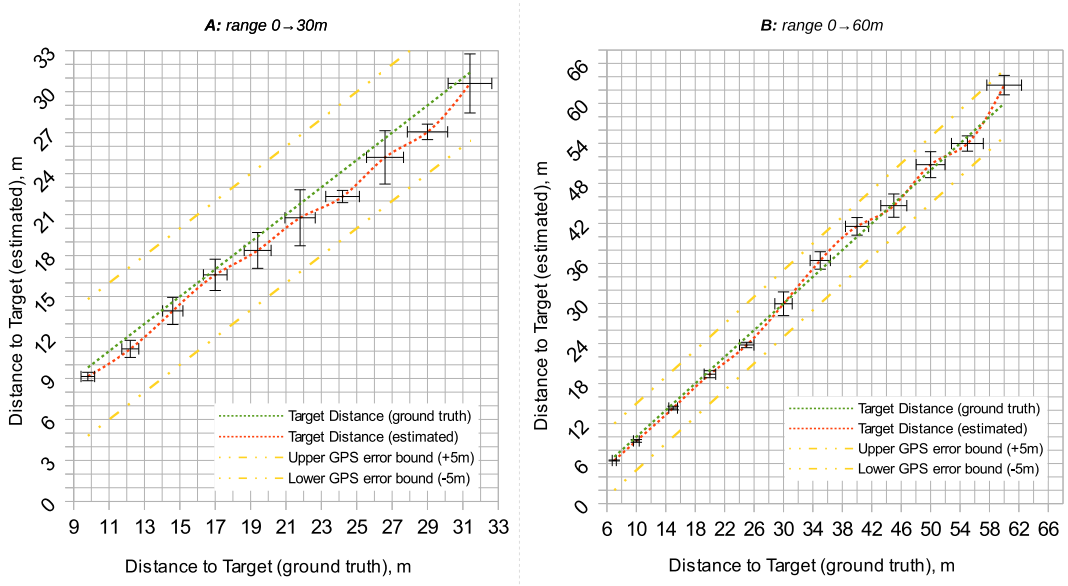

Figure 6. Estimated target distance and associated error against ground truth.

position, including error bounds, is significantly within those of GPS accuracy in both cases. Furthermore, the standard deviation obtained experimentally (y-axis error bars, Figure 6) is approximately less than equal to the expected position error due to variation in human height (x-axis error bars, Figure 6) in the majority of instances.

An excerpt from the experimental results of Fig. 6A is additionally shown in Figure 7 where we see three exemplar tracking results for a given pedestrian obtained via the photogrammetric approach outlined (Figure 7 A-C, blue track) and additionally using a hand-held GPS tracking unit (Figure 7 A-C, red track with $\pm 5m$ GPS error margin). Again we can see that the GPS tracks recovered via photogrammetry are within the spatial error margins of the GPS ground truth (in both planar axes). It is notable that the error within the X position (relative to the camera) appears larger than Z in several places. The limited sampling frequency of the hand-held GPS unit used for ground truth (1 reading per second) also limits the granularity of the GPS track (Figure 7 A-C, red) against that obtained via the camera (Figure 7 A-C, blue). As a result, some level of granularity present within the observed human track is not present within the corresponding GPS ground truth (e.g. Figure 7B).

Figures 8 - 10 present qualitative result for the detection, classification and 3D tracking of pedestrian targets within a cluttered scene environment. Within each sub-figure (Figures 8 - 10 A-G) we present the detected pedestrian(s), with associated 2D image projection of the track (right) and the planar view of the $\{Y/Z\}$ tracked position relative to the camera (left). Under test conditions, the detection of pedestrians operates with statistical accuracy of 0.91 (precision = 0.92, recall = 0.93) based on a 93% true positive and 13% false positive detection rate on a per sample basis. The use of spatio-temporal filtering within the tracking framework presented (Section 2.3.1, using $w = 10$) reduce false positives to a negligible level and produce true positive detection close to 100% based on the episodic evaluation paradigm introduced in [4, 45] (i.e. targets correctly/incorrectly detected per episode or scenario rather than per image sample).

Figure 8 A-G presents extracts from a simple pedestrian tracking example (right) and illustrates the recovery of a clear target trajectory in scene space co-ordinates (left) over a range of ∼10-35m range. Figure 9 A-G presents a recovery of a more complex, self-intersecting target trajectory in scene space co-ordinates (left) for a pedestrian tracking example (right) over a ∼10-25m range. Figure 10 presents two separate scenarios (A-D / E-G) of multiple pedestrian tracking, over a ∼10-50+m range, with recovery of more complex and intersecting target trajectories. Figure 10 E-G illustrates the disambiguation of target tracks that intersect in 2D image space (right) within the tracked 3D scene space co-ordinates (left). This disambiguation is performed based solely on the target position recovered via photogrammetry negating the need for

Figure 7. Three exemplar tracks (A,B,C) plotted on geo-referenced satellite imagery (blue) with corresponding ground truth obtained from hand-held GPS tracking units (red with $\pm 5m$ error boundary plotted).

complex tracking strategies as found in contemporary work [1, 41]. By contrast to the growing complexity of many target tracking algorithms, the proposed pipeline that enables pedestrian tracking within scene space coordinates achieves such disambiguation based on conventional Kalman filter driven tracking.

4. CONCLUSIONS

Overall we have shown that the use of photogrammetry provides an effective means for the 3D localization and tracking of pedestrians within infrared imagery. This is based on the improved pedestrian localization afforded by the use of imagery within this spectral band and the use of a background model driven detection strategy that provides tight image-space bounds on pedestrian scene regions. We have shown that the use of human height as an *a priori* constraint for pedestrian localization introduces a statistical error due to variation that is within the bounds of conventional GPS localization error for approximately 96% of the adult population for ranges up to 60m. Variance introduced due to pre-adult height variation, standard deviation across the population and statistical variation is also accounted for within our extended discussion. This is supported by experimental results that show the error to be within these bounds over a range of scenarios.

Pedestrian tracking within 3D scene-space coordinates facilitates the ready disambiguation of multiple target tracking scenarios using low-complexity approaches with reduced computational overheads. This is inherent within the premise that whilst two pedestrian targets may occupy the same position within a 2D image projection of the scene, they cannot physically occupy exactly the same 3D scene space. Our results are illustrated using a bag of visual words driven approach with spatio-temporal target detection/reporting filtering based on the earlier work of [2]. The pedestrian localization analysis extends the prior work of [4, 11] in terms of both its statistical underpinning from [5, 6, 40] and experimental error validation.

Future work will look to investigate the extension of this work to visible-band imagery, using recent advances in real-time salient object detection [46] and use within the context of mobile platform navigation [47, 48] and for multi-platform, multi-modal wide-area search and surveillance tasks [4, 49].

Acknowledgments: *This work was supported by the Defence Science and Technology Laboratory (UK MOD) and the UK Technology Strategy Board (TSB).*

REFERENCES

1. A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual Tracking: An

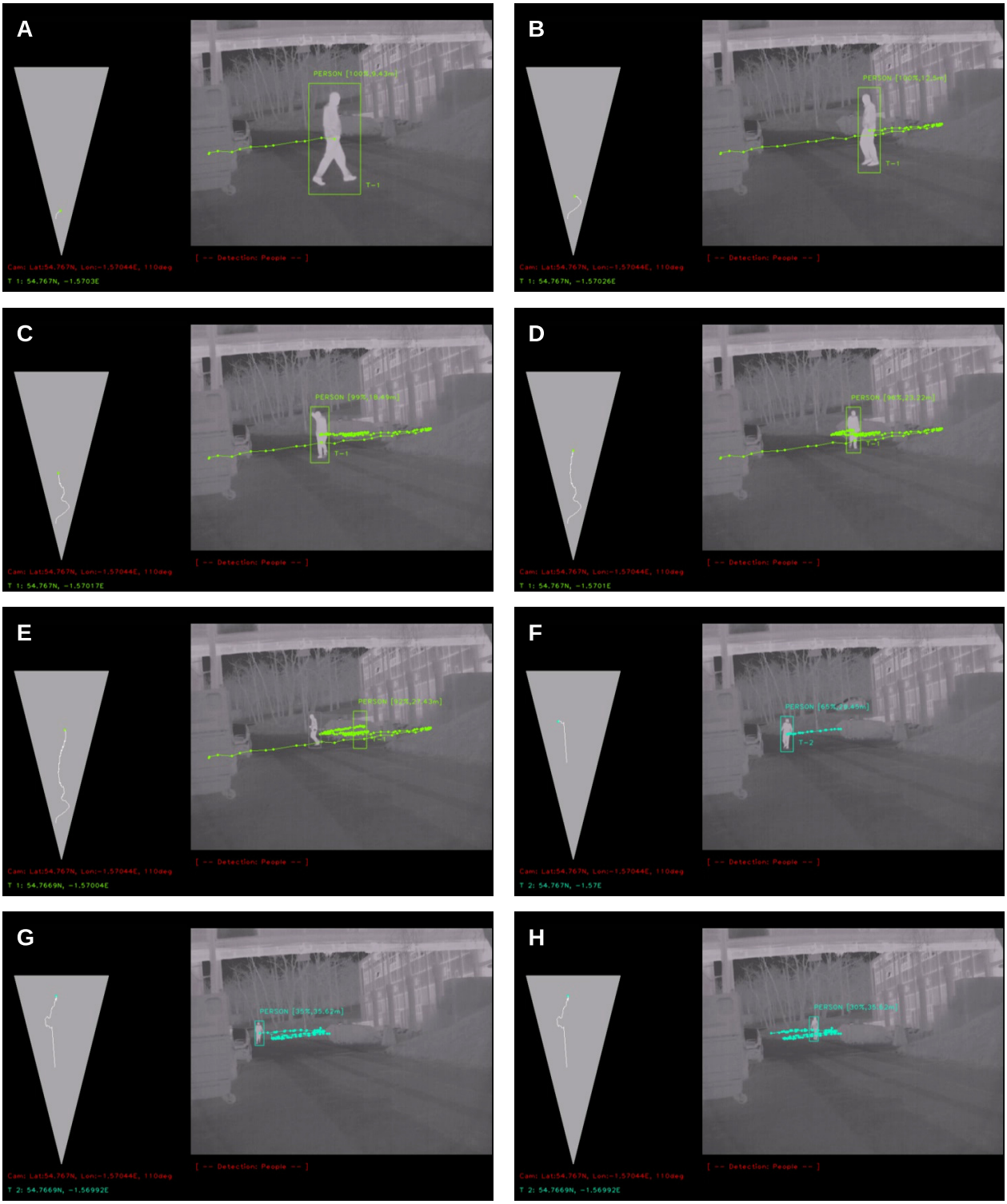

Figure 8. An example of real-time pedestrian detection and tracking in infrared imagery (right) with associated geo-referenced 3D track (left).

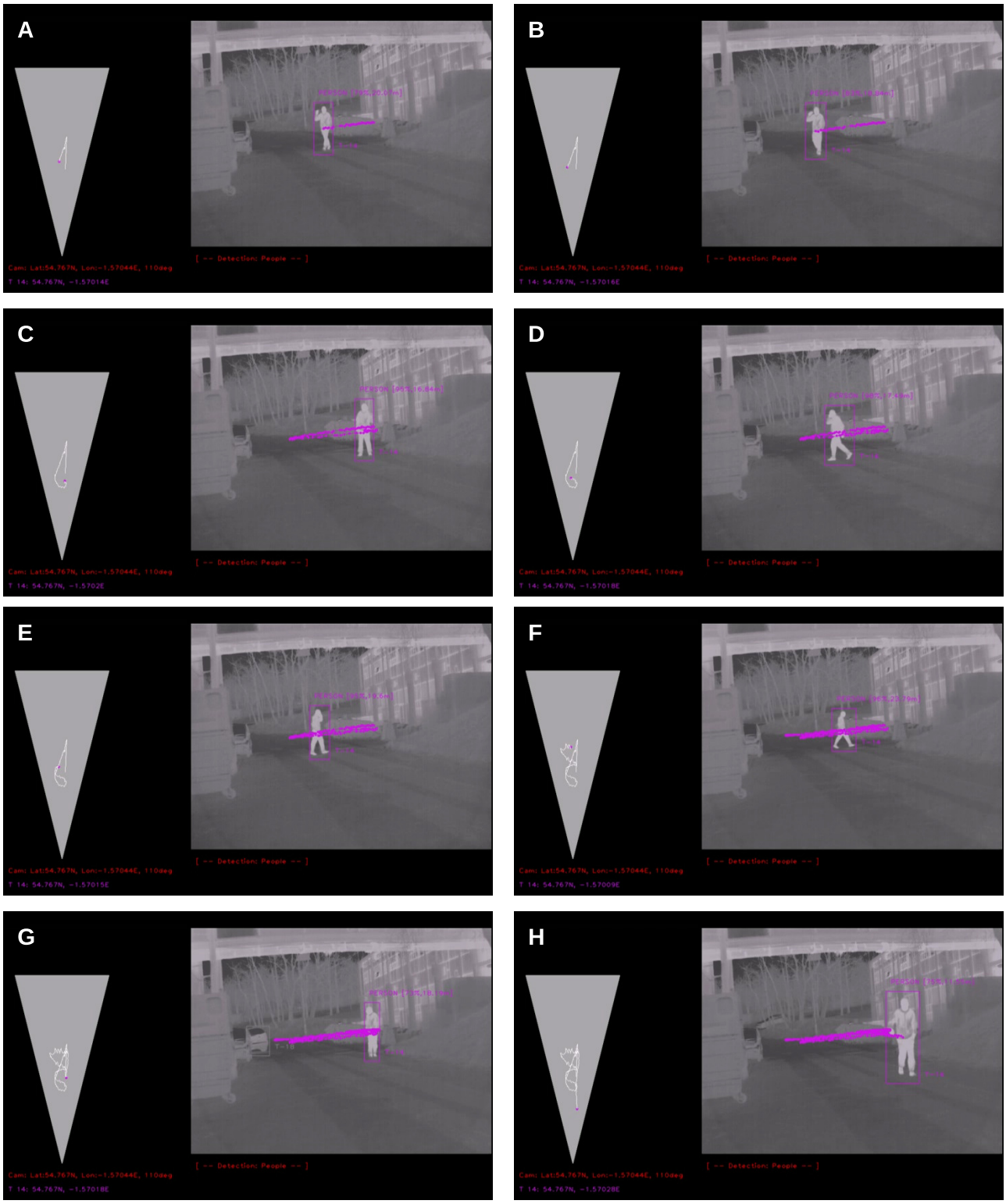

Figure 9. An example of real-time pedestrian detection and tracking in infrared imagery (right) with associated geo-referenced 3D track (left).

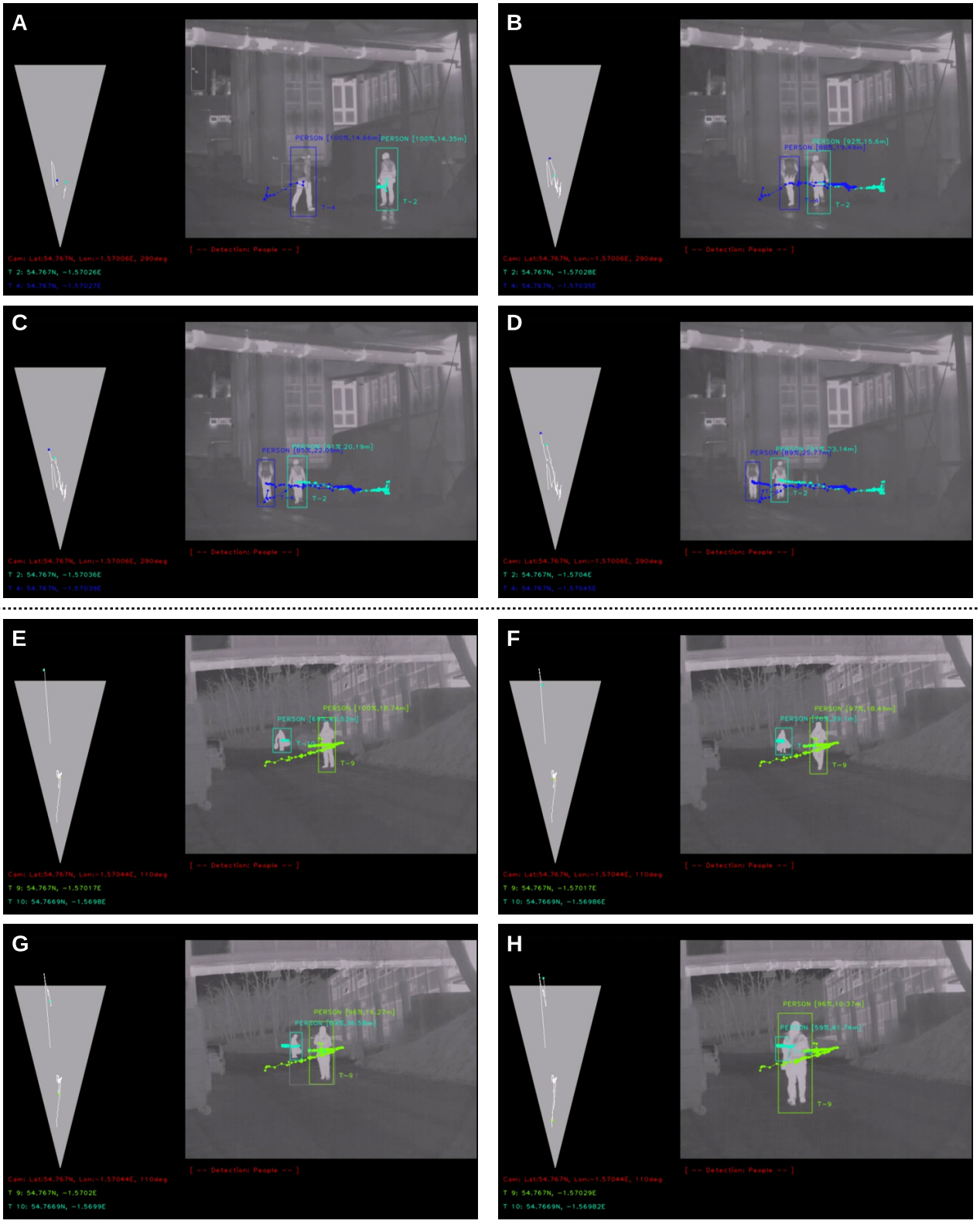

Figure 10. An example of real-time multiple pedestrian detection and tracking in infrared imagery (right) with associated geo-referenced 3D tracks (left).

Experimental Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1442–1468, July 2014.
2. T. Breckon, J. Han, and J. Richardson, "Consistency in muti-modal automated target detection using temporally filtered reporting," in *Proc. SPIE Electro-Optical Remote Sensing, Photonic Technologies, and Applications VI*, vol. 8542, pp. 23:1–23:12, November 2012.
3. J. Han, A. Gaszczak, R. Maciol, S. Barnes, and T. Breckon, "Human pose classification within the context of near-ir imagery tracking," in *Proc. SPIE Optics and Photonics for Counterterrorism, Crime Fighting and Defence*, vol. 8901, pp. 1–10, SPIE, September 2013.
4. 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*, vol. 8899, pp. 1–19, SPIE, September 2013.
5. R. Craig, J. Mindell, and V. Hirani, "Health survey for England," *Obesity and Other Risk Factors in Children. The Information Centre*, vol. 2, 2006.
6. P. M. Visscher, "Sizing up human height variation.," *Nature genetics*, vol. 40, pp. 489–90, May 2008.
7. A. Yilmaz, O. Javed, and M. Shah, "Object tracking - a survey," *ACM Computing Surveys*, vol. 38, pp. 13–es, Dec. 2006.
8. H. K. Aghajan and A. Cavallaro, *Multi-camera networks: principles and applications*. Academic press, 2009.
9. G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *J. of Ambient Intelligence and Humanized Comp.*, vol. 2, no. 2, pp. 127–151, 2011.
10. X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, 2012.
11. E. Brau, J. Guan, K. Simek, L. D. Pero, C. R. Dawson, and K. Barnard, "Bayesian 3D Tracking from Monocular Video," in *Int. Conf. Computer Vision*, pp. 3368–3375, 2013.
12. J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi, "Photo clip art," in *ACM SIGGRAPH*, vol. 26, p. 3, ACM Press, Aug. 2007.
13. J. Yuen, B. Russell, and A. Torralba, "LabelMe video: Building a video database with human annotations," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 1451–1458, IEEE, Sept. 2009.
14. G. Payen de La Garanderie and T. Breckon, "Improved depth recovery in consumer depth cameras via disparity space fusion within cross-spectral stereo," in *Proc. British Machine Vision Conference*, pp. 417.1–417.12, September 2014.
15. M. Lhuillier, "Incremental Fusion of Structure-from-Motion and GPS using Constrained Bundle Adjustments.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, July 2012.
16. D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," in *Int. Conf. Computer Vision*, vol. 1, pp. 654–661, IEEE, 2005.
17. Y. Diskin and V. K. Asari, "Dense 3D point-cloud model using optical flow for a monocular reconstruction system," in *Applied Imagery Pattern Recognition Workshop*, pp. 1–6, IEEE, Oct. 2013.
18. J. W. Davis and V. Sharma, "Robust detection of people in thermal imagery," in *Proc. Int. Conf. Pattern Recognition*, vol. 4, pp. 713–716, 2004.
19. J. W. Davis and V. Sharma, "Background-subtraction in thermal imagery using contour saliency," *Int. Journal of Computer Vision*, vol. 71, no. 2, pp. 161–181, 2007.
20. B. Besbes, A. Rogozan, and A. Bensrhair, "Pedestrian recognition based on hierarchical codebook of SURF features in visible and infrared images," in *Proc. Intelligent Vehicles Symp.*, pp. 156–161, IEEE, June 2010.
21. M. Yasuno, S. Ryousuke, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images," in *Proc. Int. Conf. Intelligent Transportation Systems*, pp. 182–187, 2005.
22. J. Wang, D. Chen, H. Chen, and J. Yang, "On pedestrian detection and tracking in infrared videos," *Pattern Recognition Letters*, vol. 33, pp. 775–785, Apr. 2012.
23. A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, May 2014.
24. P. Ricaurte, C. Chilán, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa, "Feature point descriptors: infrared and visible spectra.," *Sensors*, vol. 14, pp. 3690–701, Jan. 2014.
25. P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," in *Proc. British Machine Vision Conference*, pp. 526.1–526.12, September 2012.
26. F. Barrera, F. Lumbreras, and A. D. Sappa, "Multispectral piecewise planar stereo using Manhattan-world assumption," *Pattern Recognition Letters*, vol. 34, pp. 52–61, Jan. 2013.
27. M. G. Wing, A. Eklund, and L. D. Kellogg, "Consumer-Grade Global Positioning System (GPS) Accuracy and Reliability," *Journal of Forestry*, vol. 103, no. 4, p. 5, 2005.
28. Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.
29. A. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *American Control Conference*, pp. 4305–4312, IEEE, 2012.
30. D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R. B. Fisher, J. S. Victor, and J. L. Crowley, "Comparison of target

detection algorithms using adaptive background models," in *Proc. Int. W'shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 113–120, 2005.

31. C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. Wiley-Blackwell, 2010. ISBN-13: 978-0470844731.
32. L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Computer Vision and Pattern Recognition*, vol. 2, pp. 524–531, IEEE, 2005.
33. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering Object Categories in Image Collections," in *Proceedings of the International Conference on Computer Vision*, 2005.
34. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," in *Int. Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.
35. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A Comparison of Affine Region Detectors," *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.
36. T. Tuytelaars and K. Mikolajczyk, "Local Invariant Feature Detectors: A Survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, pp. 177–280, Jan. 2007.
37. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
38. C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
39. B. P. McEvoy and P. M. Visscher, "Genetics of human height.," *Economics and human biology*, vol. 7, pp. 294–306, Dec. 2009.
40. J. V. Freeman, T. J. Cole, S. Chinn, P. R. Jones, E. M. White, and M. A. Preece, "Cross sectional stature and weight reference curves for the UK, 1990.," *Archives of Disease in Childhood*, vol. 73, pp. 17–24, July 1995.
41. E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. Wiley, 2011.
42. J. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," tech. rep., Intel Corporation, 2001.
43. X. Li and T. Breckon, "Combining motion segmentation and feature based tracking for object classification and anomaly detection," in *Proc. 4th European Conference on Visual Media Production*, pp. I–6, IET, November 2007.
44. E. Maggio and A. Cavallaro, "Accurate appearance-based Bayesian tracking for maneuvering targets," *Computer Vision and Image Understanding*, vol. 113, pp. 544–555, Apr. 2009.
45. 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*, p. Vol. 7878 Number 78780B, Jan. 2011.
46. I. Katramados and T. Breckon, "Real-time visual saliency by division of gaussians," in *Proc. International Conference on Image Processing*, pp. 1741–1744, IEEE, September 2011.
47. I. Katramados, S. Crumpler, and T. Breckon, "Real-time traversable surface detection by colour space fusion and temporal analysis," in *Proc. International Conference on Computer Vision Systems*, vol. 5815 of *Lecture Notes in Computer Science*, pp. 265–274, Springer, 2009.
48. R. Chereau and T. Breckon, "Robust motion filtering as an enabler to video stabilization for a tele-operated mobile robot," in *Proc. SPIE Electro-Optical Remote Sensing, Photonic Technologies, and Applications VII*, vol. 8897, pp. 1–17, SPIE, September 2013.
49. M. Magnabosco and T. Breckon, "Cross-spectral visual Simultaneous Localization And Mapping (SLAM) with sensor handover," *Robotics and Autonomous Systems*, vol. 63, pp. 195–208, February 2013.