Denman, Simon and Lamb, Todd and Fookes, Clinton B. and Chandran, Vinod and Sridharan, Sridha (2008) ***Multi-spectral fusion for surveillance systems***. Computers and Electrical Engineering.

Multi-Spectral Fusion for Surveillance Systems

Simon Denman, Todd Lamb, Clinton Fookes, Vinod Chandran, Sridha Sridharan

Image and Video Research Laboratory, Queensland University of Technology, GPO Box 2434, Brisbane 4001, Australia

Abstract

Surveillance systems such as object tracking and abandoned object detection systems typically rely on a single modality of colour video for their input. These systems work well in controlled conditions but often fail when low lighting, shadowing, smoke, dust or unstable backgrounds are present, or when the objects of interest are a similar colour to the background. Thermal images are not affected by lighting changes or shadowing, and are not overtly affected by smoke, dust or unstable backgrounds. However, thermal images lack colour information which makes distinguishing between different people or objects of interest within the same scene difficult.

By using modalities from both the visible and thermal infra-red spectra, we are able to obtain more information from a scene and overcome the problems associated with using either modality individually. We evaluate four approaches for fusing visual and thermal images for use in a person tracking system (two early fusion methods, one mid fusion and one late fusion method), in order to determine the most appropriate method for fusing multiple modalities. We also evaluate two of these approaches for use in abandoned object detection, and propose an abandoned object detection routine that utilises multiple modalities. To aid in the tracking and fusion of the modalities we propose a modified condensation filter that can dynamically change the particle count and features used according to the needs of the system.

We compare tracking and abandoned object detection performance for the proposed fusion schemes and the visual and thermal domains on their own. Testing is conducted using the OTCBVS database to evaluate object tracking, and data captured in-house to evaluate the abandoned object detection. Our results show that significant improvement can be achieved, and that a middle fusion scheme is most effective.

Key words: Object Tracking, Surveillance, Thermal Imaging, Abandoned Object Detection, Particle Filter

Preprint submitted to Elsevier *26 June 2008*

1 Introduction

Surveillance and tracking systems typically use a single modality of colour video (in the visible spectrum) for their input. These systems work well in controlled conditions but often fail with low lighting, shadowing, smoke, dust, unstable backgrounds or when the foreground object is of similar colouring to the background. These conditions result in poor motion detection as well as poor tracking of an object. With advances in technology and manufacturing techniques, the cost of sensors that allow us to see into the thermal infrared spectrum has become much more affordable. Using modalities from both the visible and thermal infrared spectra, allows us to obtain more information from a scene and overcome the problems associated with using visible light only for surveillance and tracking. Thermal images are not affected by lighting or shadowing and are not overtly affected by smoke, dust or unstable backgrounds. Also, an object of interest is unlikely to be the same colour and temperature as the background. Thermal sensors on their own however are more sensitive to noise than colour sensors, and do not allow the same level of discrimination between different tracked objects (i.e. in a thermal image, two people and their clothing appear very similar due to them being close to the same temperature). Fusing the colour and thermal modalities can prove very useful in a number of applications and situations where a robust surveillance and tracking system is needed.

Tracking system traditionally have relied on a single modality as input. Tracking algorithms such as those proposed by Zhao et al [1] and Haritoglu et al [2] rely on single visible modalities, and use motion detection to locate objects for tracking. Latecki et al[3] proposed a method adapted for detection and tracking in infrared videos. A spatio-temporal representation was used, to provide a more robust method of motion detection to counter the increased noise present in IR imagery compared to visual.

Many recent tracking system have used particle filters to aid in the tracking process, as they have proved to be effective at handling occlusions and complex situations. Isard et al [4] developed BraMBLe, a Bayesian multiple-blob tracker. A multi-blob likelihood function is used to express the likelihood of a particular configuration of objects resulting in the observed image. This enabled the system to function with an unknown, time varying number of objects, allowing the tracking of multiple objects. Vermaak et al [5] proposed a Mixture Particle Filter, which addressed the problem caused by a multi-modal posterior distribution (due to ambiguities or multiple targets) causing

Email addresses: s.denman@qut.edu.au (Simon Denman,), c.fookes@qut.edu.au (Clinton Fookes,), v.chandran@qut.edu.au (Vinod Chandran,), s.sridharan@qut.edu.au (Sridha Sridharan).

poor performance. Each mode is modeled by its own particle filter, which forms part of the overall mixture, and the individual filters only interact through the computation of the weights. This overcomes problems associated with previous multi-target trackers where the samples for a given target could become deleted and the target lost. However the system still maintains just a single particle filter for the whole system, rather than one for each tracked object. Okuma et al [6] proposed the Boosted Particle Filter (BPF), and extension of [5] and used a cascaded adaboost [7] algorithm to detect the target objects to guide the particle filter. The adaboost results were also incorporated into the proposal distribution, so that when the adaboost detection performed well, the BPF distribution could incorporate this information.

Abandoned object detection (AOD) systems are often incorporated into person tracking systems [8][9][10][11]. This allows the owner of any abandoned object to be detected and tracked in addition to the abandoned object itself. It also aids in the abandoned object detection as motion that has been detected as people can be excluded from the AOD. AOD is commonly performed by using a motion detection procedure to detect medium-long term changes in a scene (i.e. a pixel was has changed from the background state but is otherwise constant)[12][13][14][15]. Spengler et al [8] proposed a person tracker and a blob based detection system to locate abandoned objects. After person tracking is performed, remaining unexplained foreground regions are extracted. These candidates are observed for a short period of time (1-5 seconds) to filter out spurious objects, and detect abandoned objects. Guler et al [9] combined a moving object detector and a stationary object detector (both based on foreground segmentation results) to locate abandoned objects and their owners. The moving object detector analyses tracked objects for splits to try and identify the drop-off events, and the resultant objects are matched against those detected by the stationary object detector. Other approaches such as [10][11] have been designed to work in a multi-camera environment, which can aid tracking and reduce the effect of occlusions. These systems move early processing results (motion detection [10], object detection [11]) the camera networks ground plane, where tracking and AOD is performed.

A few previous studies have experimented with multi-spectral fusion for surveillance and tracking. Conaire et al [16][17][18] have experimented with fusion for object segmentation, background modeling and tracking using colour and thermal infrared images. Fusion for tracking is done in the appearance model by using a multi-dimensional Gaussian to represent each pixel. The scores from the visible and thermal spectra in the appearance model are fused in different ways to match the model to the incoming image. The ways of combining scores methods are compared to ascertain the best method for this form of fusion. Some of these methods for fusion in the appearance model have been implemented in this system. Blum and Liu [19] discuss different methods of early image fusion using the wavelet transform and the pyramid transform.

These early fusion methods can be used to fuse the images before they are fed into a tracking system. Han and Bhanu [20] discuss techniques for the use of colour and infrared images in moving human silhouette extraction, as well using these silhouettes for automatic image registration between the infrared and colour images.

In this paper we aim to investigate the most effective method for fusing visual and thermal images for person tracking and abandoned object detection. We propose a modified condensation filter [21] to track and aid in the fusion of the modalities. We compare the performance of four fusion schemes for object tracking, with the performance of the visual and thermal domains on their own, and demonstrate that improvements can be achieved by using multiple modalities. We also propose and evaluate a multi-spectral abandoned object detection system. We evaluate two fusion schemes for abandoned object detection, and compare the performance of the proposed multi-spectral systems with the colour and thermal modalities individually, and show that improvements in performance can be achieved for this task as well. Section 2 will discuss the tracking system and condensation filter [21] used; Section 3 will present the proposed fusion schemes for tracking; the proposed multi-spectral abandoned object detection system is presented in Section 4; results are presented in Section 5 and conclusions in Section 6.

2 Object Tracking System

We have modified the tracking system proposed in [22] to work in a multi-modal environment. The object tracking system uses a hybrid motion detector-optical flow technique[23] that is also capable of detecting multiple layers of motion [24] as a basis. The system scans the motion detector output for appropriate regions of motion to detect people (see figure 1). A modified condensation filter (see section 2.1) is used to track the people.

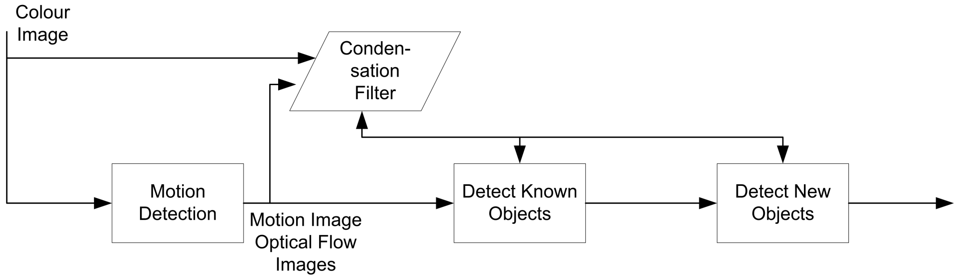


Fig. 1. Tracking System Flowchart

The condensation filter uses the colour image and the results of the motion detection (motion mask as well as optical flow) to determine the most likely positions for any known tracked objects in the current frame. This information is used to guide the person detection routines, and help evaluate matches between detected objects and tracked objects. The system uses person detection results where possible to determine the location of tracked objects, rather than relying solely on the output of the condensation filter, except in the event of occlusions. Remaining reliant on person detection results allows successful detections to be used to update the features that are in use by the condensation filter. As we are tracking people in an outdoor scene, it is likely that their appearance will change over time, and so it is important to use a dynamic model within the condensation filter, rather than one which is created when the object is initially detected and left as is for the duration of the track.

Person detection is performed by analysis of the motion image. The image is split into several sub-regions which contain concentrated areas of motion. Heads are located within each sub-region, and ellipses are fitted at the detected heads [2][1]. Working within subregions overcomes problems caused by people occupying a common column of the image causing inaccurate vertical projections. Heads are detected by combining the vertical projection and pixel height of the top contour (to aid in overcoming problems caused by holes in the motion image), and finding local maxima; which are then filtered by analysing the surrounding region. Ellipses are fitted to the valid heads at an aspect dependent on the candidate head, and if there is a suitable occupancy (motion within the bounds of the ellipse) the candidate is accepted.

When a person is detected, motion associated with the detected person is and removed from the motion image as it is now accounted for. The remaining motion must belong to new people, and so person detection is carried out on remaining areas to locate people who have recently entered the scene.

2.1 Scalable Condensation Filter

A condensation filter[21] is used to track objects in the system. We propose the Scalable Condensation Filter (SCF), an extension of the Mixture Particle Filter (MPF)[5] and Boosted Particle Filter (BPF)[6]. The SCF allows the particle count of the filter, and the features used to tracked the objects, to dynamically change. A single filter is used for the entire system, and the particle count is scaled according to the number of objects being tracked. In addition, we allow the number of the particles for each track to vary according to the complexity of the surrounding area (see section 2.1.1). As in [5][6], each tracked object's particles are resampled separately to ensure that the particles of each track (and thus the track itself) are not lost due to resampling.

Particles are four dimensional, and describe a bounding box (a centre position (x and y pixel coordinates) and the height and width, $\{x, y, h, w\}$). Each variable is free to move within the dimension limits, $\{d_{min}, d_{max}\}$, which are defined by the system (i.e. the limits of x and y are governed by the image size) and depend on the dataset being used and expected size of the objects being tracked. The distribution of each dimension is Gaussian, with the mean at the the last observed position, and the variance equal to the maximum expected movement of a dimension from one frame to the next, e_{max}.

A Sequential Importance Resampling (SIR)[25][26] procedure is used to update the sample set. Each new particle is adjusted according to a motion model associated with the tracked object responsible for the particle. The expected movement according to this motion model (based on a window of Q previous observations) is added to the particle as well as a noise vector.

$$S_{(i,n,t+1)} = S_{(i,n,t)} + M_i + R \tag{1}$$

where $S_{(i,n,t+1)}$ is the nth sample for track i at the next time step; $S_{(i,n,t)}$ is the nth sample for track i at the current time step; R is the random sample, which is within the range of $-e_{max}$ to $+e_{max}$, and M_i is the expected movement for the track, i. As part of all particle updating and creation, a set of limits is applied to each particle, to ensure that it is describes a valid object (if a dimension exceeds a limit, it is set to the limit). Whilst SIR would ensure that any particles that describe invalid objects are not propagated (they would have 0 probability), performing this test on the particles at this point avoids the need to check for valid image coordinates when matching features, and allows fewer particle to be used as all are guaranteed to valid. This allows the system to be more efficient.

2.1.1 Dynamic Sizing

Rather than have a fixed number of samples for the filter, the sample count is dynamically changed as objects enter and leave the scene, and as objects move about and occlude one another. For each track, an arbitrary number of samples, n, are created about the objects initial position and associated with that object.

$$s_{new} = o_{new} + 2 \times r \tag{2}$$

where s_{new} is the new sample, o_{new} is the new objects state, and r is a random value, in the range $-e_{max}$ to $+e_{max}$.

The particles initially associated with the given track remain associated with it for the duration of the track's life. This initialisation gives each tracked object a set of samples to model it immediately, rather than needing to allow a period of frames for the system to adapt to its presence. When an object leaves, all particles belonging to the track are removed from the system.

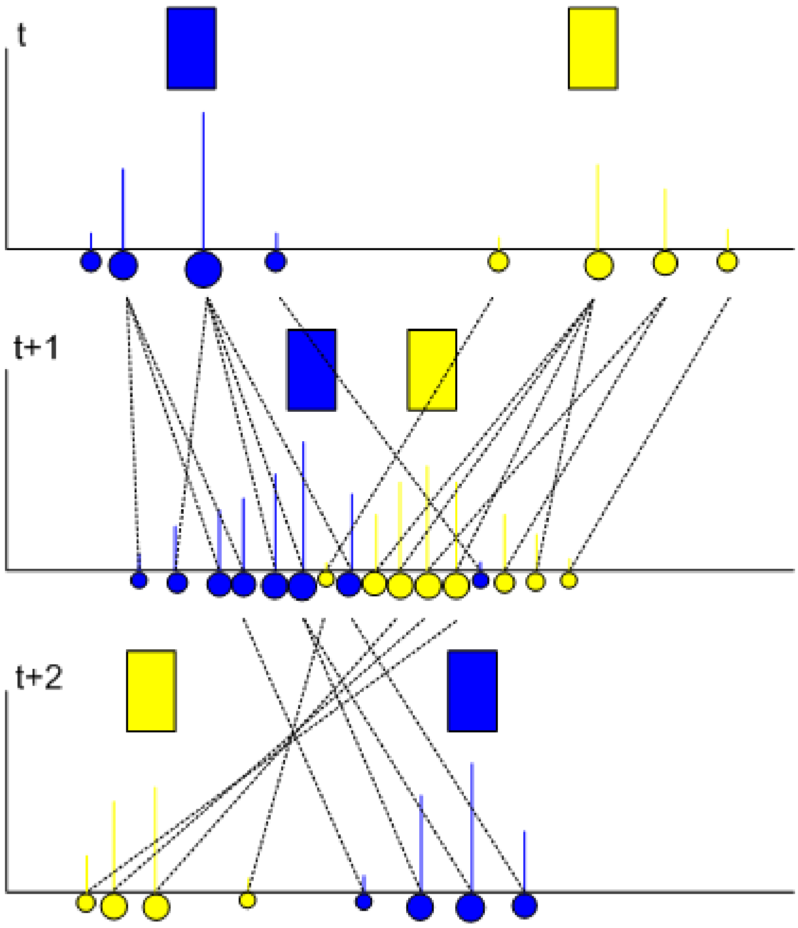


Fig. 2. Dynamic Sizing of Particle Filter - At time t, the two tracked objects are suitably far apart that there is no occlusion, and so each object is tracked with the standard number of particles (4). At time $t+1$, the objects are considered to be in an occlusion state. The system oversamples the particle set to generate a set of 8 particles for each of the tracked objects. At time $t+2$, the occlusion has passed, and so the sample sets are undersampled such that each object is once again tracked by the standard number of particles.

When two or more tracked objects are close together, additional particles can be added and more advanced features can be used to aid in the tracking. Three levels of occlusion are defined with the system for each track:

(1) Level 0 (No Occlusion) - The tracked object is isolated within the scene, there are no other objects nearby
(2) Level 1 (Object Nearby) - Another tracked object's bounding box is within a distance r
(3) Level 2 (Overlap) - Another tracked object's bounding box is overlapping

When a track is first created, and added to the SCF, it is at occlusion level 0 and is created with the standard number of particles. For each occlusion level increase, an additional m particles are added to the SCF for that track; and m samples are removed for each occlusion level decrease. Particle counts for tracked objects are altered during the resampling procedure by either undersampling or oversampling (see Figure 2).

Resizing the system in this manner ensures that no unnecessary updates are done, and improves CPU utilisation.

2.1.2 Object Features

Each track is able to use multiple features. Using inheritance and polymorphism, the types of features used by each track can vary depending on the circumstances (i.e. occlusion) and the class of object being tracked, without any change required in the condensation filter itself. This approach allows different types of objects to use feature more suited to their individual properties (i.e. a person and a car can be tracked by the same condensation filter, yet use different appearance models that better represent the target objects).

Two classes of features are defined, each of which has various sub-types.

(1) Histograms
(2) Appearance Models

Histograms simply model colour distributions, and so while being quicker to compute, do not take structural information into consideration (i.e. a person wearing blue pants and a red shirt will have a very similar histogram to a person wearing red pants and a blue shirt, despite having a distinct appearance). Appearance models encode position information as well as colour information, and so are more discriminative. They are however more processor intensive. Each of these features can optionally use motion detection and optical flow as additional aids (i.e. a pixel must be in motion and must be moving in the same direction as the object being tracked), and the use of these cues can change dynamically depending on the systems status (i.e. if motion detection is unreliable for a period of time due to environmental effects, this can be omitted when matching features).

The features used by the system vary as the complexity of the system changes. A histogram feature is used by default, and when a track's occlusion level increases above 0 (see Section 2.1.1) an appearance model feature is used as well. When multiple features are used, the probability for the particle is the product of the probabilities for each feature. As each tracked object has its probabilities normalised, and particles resampled separately, there is no danger of the additional matching constraints reducing a track's probabilities to the extent that the track's particles are removed from the system by the resampling procedure.

2.1.3 Proposed Appearance Model

An appearance model that utilises the motion detection routine used within this tracking system is proposed. The model incorporates colour, motion state, and optical flow into a single model. The appearance model, A, is a grid of A_x by A_y squares, with an average colour ($A_c(k)$, where k is the colour channel), velocity (A_u and A_v for the horizontal and vertical velocity respectively, derived from the optical flow), and motion occupancy (A_m) stored for each square. An error value for the colour (A^e_c) and optical flow (A^e_{opf}) is also stored for each square. The appearance model is updated using equations 9, 10 and 11, and the equivalent features computed for the incoming images (i.e. the average colour, optical flow, motion occupancy, colour error and optical flow error for the input image, see Equations 3, 4, 5, 6, 7 and 8).

The input image, $I(t)$ is divided in to a grid of dimensions A_x by A_y. It is assumed that these dimensions will be significantly smaller than those of the input images (see Figure 3).

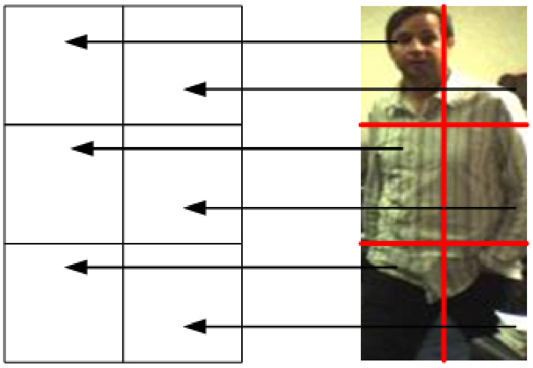


Fig. 3. Dividing input image for Appearance Model

For each grid square in $I(t)$, the average colour, percentage of motion, and optical flow (horizontal and vertical) are computed.

$$F_c(x', y', t, k) = \frac{1}{card(M(x, y, t))} \sum I(x, y, t, k) \; where \; (x, y) \in M(t), \quad (3)$$

$$F_u(x', y', t) = \frac{1}{card(M(x, y, t)}) \sum U(x, y, t) \; where \; (x, y) \in M(t), \quad (4)$$

$$F_v(x', y', t) = \frac{1}{card(M(x, y, t))} \sum V(x, y, t) \; where \; (x, y) \in M(t), \quad (5)$$

$$F_m(x', y', t) = \frac{card(M(t))}{card(I(t))} \quad (6)$$

where F is a feature extracted for the current image, x', y' are in the range $[0..A_x - 1, 0..A_y - 1]$, U and V are the input horizontal and vertical flow image, M is the input motion image and $M(t)$ is the set of all pixels that are in motion, $card(M(x, y, t))$ is the size (cardinality) of the set $M(x, y, t)$, and x, y is in the range that corresponds to the grid square x', y'.

When performing an update or comparison, an error measure is also calculated for the colour and optical flow components,

$$F_c^e(x', y', t) = \sum_1^K |A_c(x', y', t, k) - F_c(x', y', t, k)|, \quad (7)$$

$$F_{opf}^e(x', y', t) = |A_u(x', y', t) - F_u(x', y', t)| + |A_v(x', y', t) - F_v(x', y', t)|, \quad (8)$$

where F_c^e and F_{opf}^e are the frame errors for colour and optical flow respectively, and K is the number of colour channels in the appearance model.

Given the features and error measures for the incoming image, the appearance model components are updated according to the equation

$$A(t+1) = A(t) + (F(t) - A(t)) \times L, \quad (9)$$

where L is the learning rate. L is defined as

$$L = \frac{1}{T}; T < W, \quad (10)$$

$$L = \frac{1}{W}; W >= T, \quad (11)$$

where W is the number of frames used in the model, and T is the number of updates performed on the model. This ensures that the image use for model initialisation does not dominate the model for a significant number of frames. Instead, the information is incorporated quickly when the model is new to provide a better representation of the tracked object being modeled sooner.

Equation 9 is applied to each individual appearance model component and its corresponding feature from the incoming image, such that the updates for the colour and colour error components become

$$A_c(x', y', t+1, k) = A_c(x', y', t, k) + (F_c(x', y', t, k) - A_c(x', y', t, k)) \times L, \quad (12)$$

$$A_c^e(x', y', t+1) = A_c^e(x', y', t) + (F_c^e(x', y', t) - A_c^e(x', y', t)) \times L. \tag{13}$$

Updates for the other components (horizontal and vertical flow, optical flow error and motion occupancy) are performed similarly.

The errors are stored in the appearance model and updated over time using (see Equations 9, 10 and 11). The cumulative error is used as an approximation to the standard deviation (we assume that the observations over time form a Gaussian distribution) of the error, as it is not practical to re-compute the standard deviation each frame, and not ideal to assume a fixed standard deviation. Given that the standard deviation for a sample set is defined as

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mu - s_n)^2}; \tag{14}$$

and in the proposed appearance model, for each grid square there is one observation at each time step ($N = 1$), so the standard deviation at a given time step is

$$\sigma = \sqrt{(\mu - s)^2} = |A(x', y', t) - F(x', y', t)|, \tag{15}$$

which is the proposed error measure.

When matching the model to an input image, average colour, flow and motion occupancy is computed for the image in the same manner as for an update. Errors for the colour and optical flow are calculated and these are compared to the cumulative errors for the model

$$D_c(x', y', t) = \frac{A_c^e(x', y', t)}{F_c^e(x', y', t)}, \tag{16}$$

$$D_{opf}(x', y', t) = \frac{A_{opf}^e(x', y', t)}{F_{opf}^e(x', y', t)}, \tag{17}$$

where $D_c(x', y', t)$ and $D_{opf}(x', y', t)$ are the number of standard deviations from the mean that the observation (input image) is. A normal distribution look-up table is used to determine the probability that these observations have arisen from the model, which yields $P(F_c(x', y', t)|A_c(x', y', t))$ as the probability that the colour observation belongs to the distribution described in the appearance model, and $P(F_{opf}(x', y', t)|A_{opf}(x', y', t))$ as the probability that the optical flow observation belongs to the distribution described in the appearance model.

The probability that a given grid square matches the corresponding area in the input image is then defined as

$$P(F(x',y',t)|A(x',y',t)) = (F_c(x',y',t)|A_c(x',y',t)) \times (F_{opf}(x',y',t)|A_{opf}(x',y',t)), \tag{18}$$

where $F(x',y',t)$ is the set of features for a grid square in the input image, and $A(x',y',t)$ is the set of features for a given grid square in the appearance model.

The motion occupancy component of the model is used as a weight when computing the match across the whole model. A higher motion occupancy indicates that there is more motion, and thus more information, in a given grid square. Given this, the match for the model to an input image is

$$P(I(t)|A(t)) = \frac{\sum_{x'=1;y'=1}^{x'=A_x;y'=A_y} P(F(x',y',t)|A(x',y',t)) \times A_m(x'y',t)}{\sum_{x'=1;y'=1}^{x'=A_x;y'=A_y} A_m(x',y',t)}. \tag{19}$$

3 Fusion for Object Tracking

To determine the most appropriate method for fusing the thermal infrared and visible light images for object tracking, four different fusion approaches are proposed (see figure 4):

(1) Fusing images during the motion detection by interlacing the images
(2) Fusing the motion detection results of each image
(3) Fusing when updating the tracked objects using detected object lists from each modality
(4) Fusing the results of two object trackers, which each track a modality independently

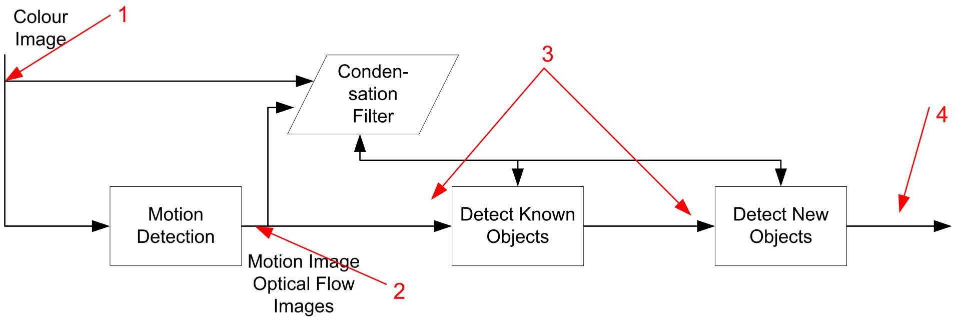


Fig. 4. The points for fusion in the system

For each of these proposed systems, the tracking system described in Section 2 is used (with any required modifications made to allow for the fusion process). In all cases, the scalable condensation filter is used to support the tracking, using a histogram model and the proposed appearance model (see Section 2.1.3).

3.1 Fusion in the Motion Detector

The first fusion method involves fusing the images prior to the motion detection by interlacing the luminance channel of the visible light image with the gray scale thermal infrared image. This approach is facilitated by using a motion detector which requires YCbCr 4:2:2 input [23]. The motion detector analyses images in 2 pixel (four value, two luminance, one blue chrominance and one red chrominance) blocks from which clusters containing two centroids (a luminance and chrominance cluster, $\{Y_1, Y_2; Cb, Cr\}$) are formed. The centroids of the clusters in the background model are compared to those in the incoming image to determine foreground/background.

Rather than convert the colour image to YCbCr 4:2:2 format as would be done in normal circumstances, it is converted to YCbCr 4:4:4. The thermal information is then interlaced with the colour information. By treating the thermal information as additional luminance data and doubling the luminance information, we effectively create a YCbCr 4:2:2 image (see figure 5) that can be fed directly into the tracking system without any further modifications. This results in the motion detector clusters becoming $\{Y, T; Cb, Cr\}$. This method of fusion has the advantage of consuming little processing resources on top of our existing system (the only additional load is when performing motion detection), and is also very simple to implement. It does however

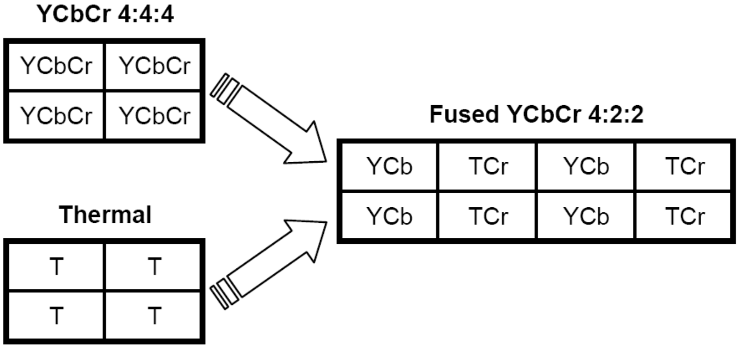


Fig. 5. Fusing Visual and Thermal Information - The YCbCr 4:4:4 representation of the input images is combined with the thermal image to produce a YCbCr 4:2:2 image where every second luminance value is actually the corresponding thermal value.

require that the colour and thermal images be correctly registered, which may require additional processing, or in some situations, not be possible.

3.2 Fusion After Motion Detection

The use of middle or late fusion allows for greater control over the information contained in the images that can be used by the tracking process. This information can be used to greatly improve the accuracy and robustness of the detection and tracking system. In both the second and third (see section 3.3) of the proposed fusion systems we compute motion detection for both images. If either image shows an abnormal increase in motion, it is disregarded. In the unlikely event that both show such an abnormality, the more consistent of the two is chosen. The abnormality of the images is assessed by looking at the percentage increase of the in motion pixel count.

$$\frac{M_t}{M_{t-1}} > T \tag{20}$$

where M is the amount of motion in the image, t is the time step and T is the threshold for determining invalid motion detection results. This test is not performed if the overall percentage of pixels in motion in the scene is beneath a threshold (10% in our system), as if there is very little motion then something such as a person entering the scene may be enough to result in an invalid image.

Our second proposed fusion scheme involves fusing directly after the motion detection. Once the motion detection masks are obtained for each the visible light and the thermal infrared modalities, they are combined to obtain a single mask for the scene. Rather than simply apply a logical "and" or "or" operation, we propose fusing the images the follow equations.

$$(M_{IR}(x,y,t) > T_1) \& (M_{Vis}(x,y,t) > T_1) \tag{21}$$

$$M_{IR}(x,y,t) > T_2 \tag{22}$$

$$M_{Vis}(x,y,t) > T_2 \tag{23}$$

where M_{IR} is the thermal motion image, M_{Vis} is the visual motion image, and T_1 and T_2 are thresholds to control the fusion ($T_2 > T_1$). If any other these equations are satisfied, the fused motion mask at (x, y, t) is set to indicate motion. The resultant mask is used in the remainder of the system described in section 2.

3.3 Fusion After Object Detection

A second mid-fusion scheme is evaluated whereby motion detection and object detection is carried out on both modalities, and the two object lists are used to update the central list of tracked objects. Objects that have been previously detected can be updated by a detection from either domain. For a new object to be added, the object must be detected in both, or in the modality where it is not detected there must be a given amount of motion within the region where the object has been detected. The amount of motion required in the second modality (where the object has not been detected) is the pixel count for the detected object multiplied by a value, T. T, for our system is 0.5. This attempts to ensure that a false detection in one modality, does not lead to an non-existent track being initialised.

The proposed appearance model (see Section 2.1.3) is extended to contain information from both motion detection routines. An additional motion and optical flow component are added, such that the model consists of a shared colour component, a motion and optical flow component for the visual domain input and a motion and optical flow component for the thermal domain. The model can be used to compare a detected object to either domain individually, or to both simultaneously (an update can also be performed on only a single domain, or both).

3.4 Fusion After Tracking

A late fusion scheme is evaluated where each modality is tracked individually and the resultant tracked object lists are fused in the same manner as for a multi-camera network. Each view is processed separately, and a list of tracked objects from each view is generated and tracked independently. At the end of each frame, a camera management module attempts to determine what objects that are being tracked by the individual trackers, represent the same real-world objects. As our multi-camera network consists of two cameras that are observing exactly the same area, there is no need to transfer to a world coordinate scheme, or rely of camera calibration, pixel coordinates can be used directly. However, as one view is in the colour domain and one is in the thermal, we are unable to use colour/appearance as an additional metric. Given this, we simply use the overlap of the bounding boxes to group objects.

At the end of each frame, the object lists are compared. It is expected that all objects should be tracked in both modalities. For those objects that are being tracked in only one modality, the tracks in the second modality (that are not already associated with a track in the first) are searched to find a match based on the overlap of the bounding box. If a matching track cannot be found (presumably due to an inability to detect the object due to poor motion detection), one is created and the system will attempt to begin tracking in the next frame (in this case, the new track is initialised without initialising histograms and appearance models, and as a result the condensation filter cannot be used until these have been initialised).

Tracked objects that have been paired across the views are compared each frame, to check that they are, in fact, a valid match. If the overlap between these two objects drops below a threshold for two consecutive frames, the pair is broken up, and at the end of the next frame the system will attempt to pair the tracks again (it is possible that they will be paired with each other again).

4 Fusion for Abandoned Object Detection

4.1 Single Modality Abandoned Object Detection System

The abandoned object detector works at a pixel level to locate abandoned objects. Individual pixels detected as being abandoned (abandoned pixels) are grouped spatially and temporally. When a grouping reaches a size threshold, an abandoned object is detected. The process builds directly on the results produced by the multi-layer motion detector [24]. The motion detection routine separates the detected motion into active motion (objects that are currently moving), and static motion (objects that are not part of the background, but are currently not moving). Multiple static motion modes (layers) are able to exist at a given pixel, allowing situations such as one car stopping in front of another stopped car, to be modeled.

The static motion image is used as input for the abandoned object detection (AOD). Abandoned pixels are created for all pixels in motion within the static motion image. Each abandoned pixel has a time stamp (for when the pixel becomes abandoned) and the pixel's colour and its layer (depth) in the static motion image stored. At each time step, the abandoned pixels are updated. An accumulator image that indicates how long a pixel has been present is constructed in the same manner as the timers that are used for the static motion image. The static layer number and colour information are used to determine matching abandoned pixels in the accumulator image. By using the static layer image, we allow multiple abandoned pixels to be present at a given location in the image, allowing overlapping abandoned objects to be detected and segmented correctly (see Figure 6).

Once an abandoned pixel's counter reaches a threshold, it is added to the

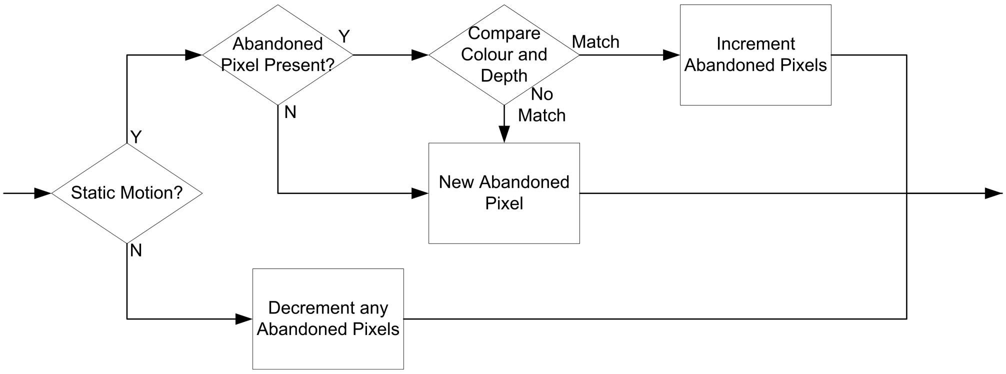


Fig. 6. Abandoned Object Detection Process for each Static Pixel - If a static pixel is present, attempt to update any existing AOD pixels that exist at this location. If a match cannot be found, or no AOD pixel exists, create a new one. If there is no static motion, decrement any abandoned pixel counters.

list of abandoned objects. The list of objects is searched for an object that is eight-connected (i.e. one or more pixels that is horizontally, vertically or diagonally adjacent to the pixel in question belongs to an abandoned object) to the newly abandoned pixel, and if such an object is found, the pixel is added to the object. If no such object can be found, a new abandoned object is created. Merging and splitting of objects occurs at the end of each processing loop, to account for newly abandoned pixel joining objects. As we are using a static motion image that is reliant on a presence threshold to add pixels, this threshold is normally kept very low, and it is simply present to add an extra level of control.

4.2 Multi-Spectral Abandoned Object Detection

The proposed multi-spectral AOD system uses a separate multi-layer motion detector for both the thermal infrared and visible light inputs. In the single spectrum AOD (see Section 4.1) the accumulator is incremented by a given value when the incoming static motion and colour is matched to a stored abandoned pixel. In the multi-spectral AOD, there are two incoming static motion images. Despite this, there is still only one accumulator used by the AOD. The AOD stores both static motion values for both the colour and thermal modalities, in addition to the colour, time stamp and counter.

Like the single modality system, incoming layer values and colours are matched to the AOD pixel. When the system matches a static motion pixel to an AOD pixel, it requires both the static layer in the thermal motion detector output, and in the visual motion output, to match those stored by the AOD pixel (unless one image is not registering static motion). If both modes match, the AOD pixel is increased by a user supplied increment (Inc_1). If only a single mode matches (i.e. one mode is failing to register static motion), then the

AOD pixel counter is increased by a separate increment (Inc_2). Depending on the requirements of the system, this increment may be same as the first, significantly less, or even negative (this may be used in a situation where there is high noise and one or both modes are unreliable). If there is no match to a static layer, all AOD static layer accumulators are decremented by a supplied value (Dec_1). The two increments and decrement are related as follows,

$$Inc_1 \geq Inc_2 \geq Dec_1. \tag{24}$$

If static motion is detected at a pixel and there is no recorded AOD previously at that pixel, or if the static motion that is detected does not match an existing AOD pixel, a new AOD pixel is created. This process is shown in Figure 7. Abandoned objects are built in the same manner as they are in the single mode system.

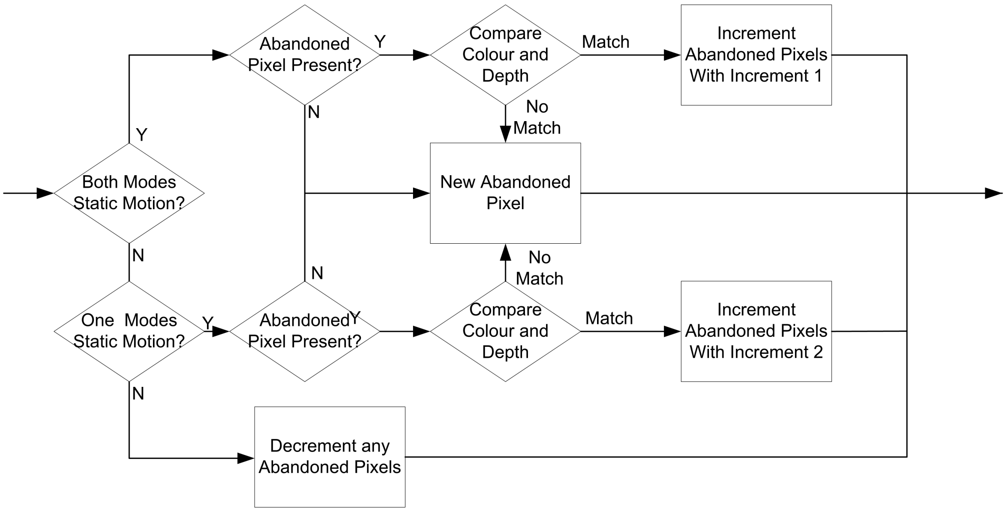


Fig. 7. Multi-Spectral AOD Flowchart

4.3 Fusion with Multi-Spectral Tracking System

The abandoned object detection is incorporated into the system as shown in Figure 8. Two fusion schemes are evaluated for this application, those detailed in Section 3.1 and Section 3.3. As multi-spectral AOD has been developed to use two motion detector images (one from each modality), the first fusion scheme is tested using the original abandoned object detector (see Section 4.1). The other fusion schemes proposed earlier are less suitable for this applications. The second proposed scheme fuses the results of the two motion detectors, and the fourth uses independent trackers (so the motion detection results from one modality are not accessible from the other).

After each frame, the motion that has not been assigned to tracked objects

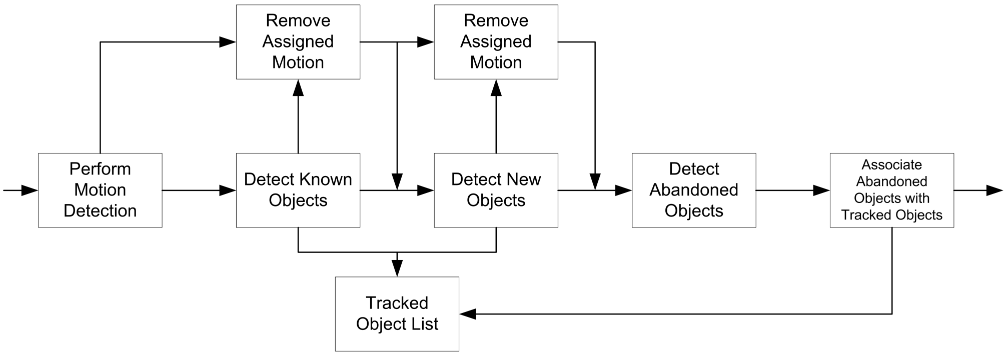


Fig. 8. Tracking System - Motion detection is used to detect objects in two stages; detect known (previously detected) objects, followed by detecting any new objects. The remaining motion (which does not belong to people) must belong to any abandoned objects and is used to update the abandoned object detector. The system then attempts to abandoned objects with their owners.

(i.e. is unaccounted for) is used to update abandoned object detector.

4.3.1 Abandoned Object Owner Matching

By integrating the abandoned object detection into a tracking system it allows the person that dropped the luggage to be detected and tracked. However, there is no guarantee that a person who drops a bag will stay near it until it is detected as an abandoned object, to allow for easy identification. In order to match an abandoned object to its owner, it is necessary to store the location information of all of the tracked people in the scene.

Timestamped locations of each tracked object are stored at the end of each frame. When an abandoned object is detected, the system then searches through the system history to find which person was in the vicinity of the abandoned object at the time in which it is calculated to have appeared. This approximate time of placement for the abandoned object can be determined by working backwards from the thresholds that are used to determine static motion layers and abandoned objects. The overlap of the tracked persons bounding box with the abandoned object is calculated and the tracked object with the greatest overlap is identified as the owner of the abandoned object.

5 Results

5.1 Object Tracking

The OTCBVS Benchmark Dataset Collection[27] is used to evaluate the four fusion tracking systems. This is a publicly available dataset that contains aligned thermal infrared and colour image sequences of two different outdoor scenes containing pedestrians. The sequences include a variety situations of interest with multiple pedestrians to test the system. We test the performance of the proposed fusion system as well as tracking with both modalities individually.

Eight sub-sequences from the database are selected to highlight various situations of interest such as stationary people, occlusions, people moving in shadowed areas, and shadowing caused by cloud cover. Three sequences from the second location (set 1 in our evaluation), and five from the first (set 2 in our evaluation) are used. Separate results are shown for each set of sequences, as the first set (taken from Location 2 in the database) contains significantly simpler scenarios than those in the second. Ground truth tracking data has been computed for each of these sub-sequences using the VIPER toolkit [28].

Tracking output is compared to the ground truth data using the ETISEO evaluation tool [29], developed as part of the ETISEO evaluation. The ETISEO evaluation defined several metrics for gauging the performs of tracking systems, which are split into five groups:

(1) Detection
(2) Localisation
(3) Tracking
(4) Classification
(5) Event Recognition

Results for the proposed tracking systems will be evaluated using metrics from the first three groups (there is only one type of object being tracked in the system, people, and there is no event recognition). Each group of metrics contains several metrics to evaluate specific areas of interest and a global metric, which is the average of the all metrics within the group. In addition to the global metrics for detection, localisation and tracking, our evaluation will also show results for the metrics defined in Table 1. All metrics result in a value in the range $[0, 1]$, with 1 being a perfect result, and 0 being complete failure. Detailed information on how the metrics are formulated can be found in [30].

Results for the first set of three sequences are shown in Table 2 and Figure 9.

Metric	Description
D1	The number of detected objects that have a significant overlap with a ground truth object. Only one detected object can match a ground truth object, any additional detections are designated as false positives (further details in [30], metric M1.2.1).
L1	Evaluate the 2D object position in each frame. Uses the overlap of the bounding ground truth and result bounding boxes (further details in [30], metric M2.1.1).
T1	Measures the percentage of time that an object is tracked for. Assumes that the object ID will be constant over the object life, and uses the distance between ground truth and result data to determine corresponding tracks (further details in [30], metric M3.3.1).

Table 1
Evaluation Metric Standard Definitions

Tracking System	D1	L1	T1	Overall Detection	Overall Localisation	Overall Tracking
Colour	0.88	0.76	0.46	0.58	0.92	0.67
Thermal	0.88	0.78	0.69	0.69	0.93	0.72
Fusion 1	0.91	0.76	0.49	0.52	0.93	0.75
Fusion 2	0.89	0.77	0.50	0.63	0.93	0.77
Fusion 3	0.94	0.83	0.71	0.73	0.95	0.84
Fusion 4	0.88	0.79	0.63	0.65	0.93	0.72

Table 2
Set 1 Results for Proposed Fusion Systems

As Table 2 shows the proposed fusion systems offer an improvement over using either modality individually. The colour modality alone performs the worst of the tested systems. This can be attributed to noise that is present in the colour dataset, that resulted in poor performance for the motion detection (when compared to the thermal modality). This in turn impacted upon the detection performance, and tracking performance. This had differing effects on the fusion systems. With the exception of the third fusion system, all fusion systems failed to outperform the thermal system in detection due to the noise that was carried through in the motion detection. The third fusion system treats both modalities separately for detection and picks the best match to use in an update. This approach results in noticeable improvements in the detection and tracking performance. Whilst the fourth system also treats both modalities separately, it also tracks the objects independently in each view. This means that a poor match may still be used to update in one modality, the effects of which may not be totally canceled by the other. Localisation performance is very similar for all systems. This can be attributed to all systems using the same person detection routines with the same parameters.

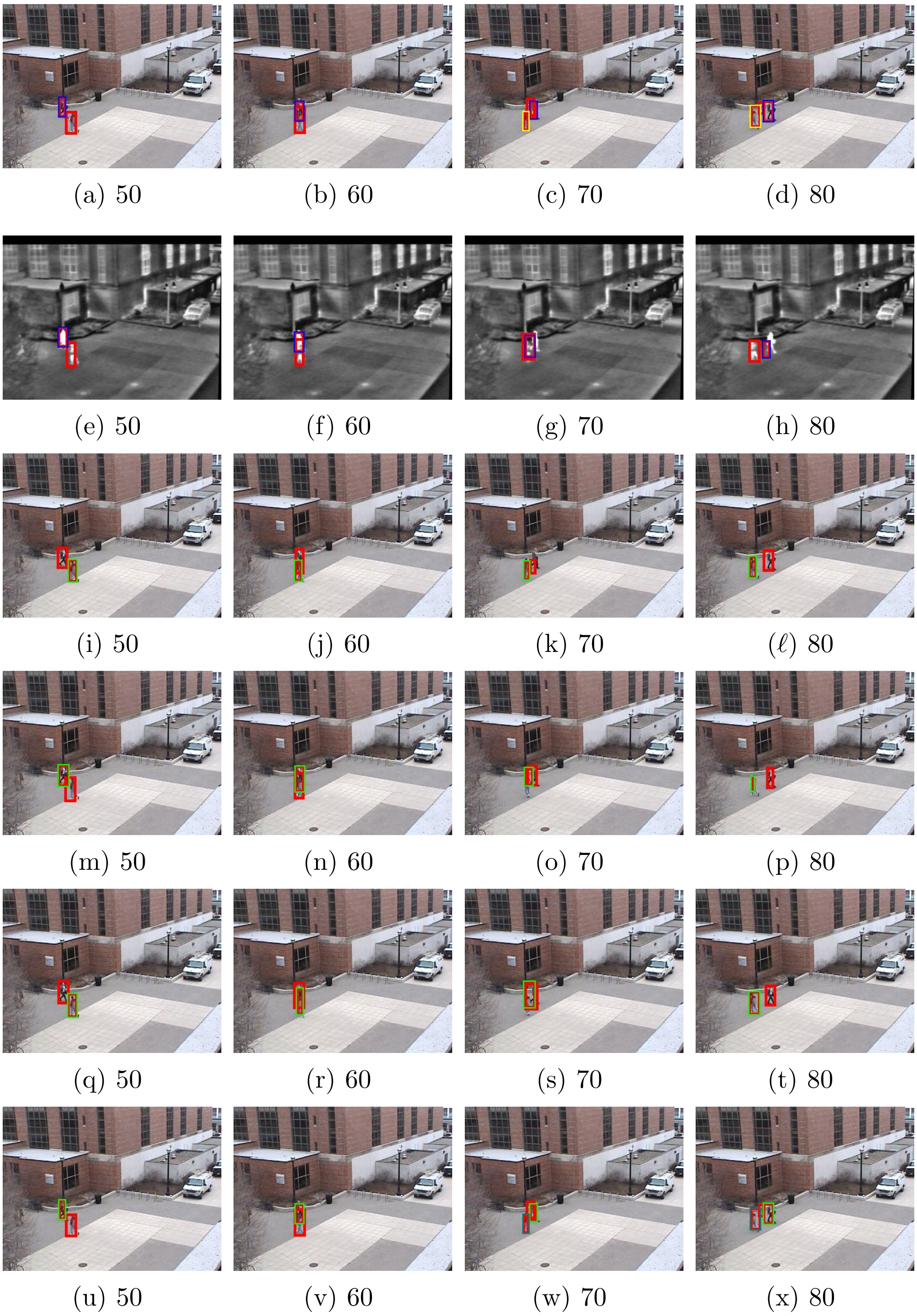
Figure 9 shows an example of the tracking output from set 1, where two people cross paths, causing an occlusion. With the exception of the colour only modality and fourth fusion system, each configuration is able to resolve the occlusion correctly (although the thermal modality does mis-track badly before correcting). The failure in the colour modality can be attributed to the poor motion and object detection performance, and also explains the failure in the fourth fusion system, as the fourth system attempts to combine results for the two modalities.

Results for the second set of five sequences are shown in Table 3 and Figures 10 and 11.

Tracking System	D1	L1	T1	Overall Detection	Overall Localisation	Overall Tracking
Colour	0.76	0.59	0.24	0.40	0.87	0.34
Thermal	0.89	0.74	0.44	0.52	0.92	0.52
Fusion 1	0.82	0.65	0.40	0.48	0.89	0.46
Fusion 2	0.70	0.60	0.28	0.39	0.88	0.35
Fusion 3	0.87	0.74	0.47	0.59	0.92	0.52
Fusion 4	0.81	0.71	0.44	0.50	0.90	0.46

Table 3
Set 2 Results for Proposed Fusion Systems

As Table 3 shows, the third proposed fusion scheme achieves the best performance, slightly ahead of the thermal modality individually. Examples of the system output are shown in Figures 10 (simple scenario with several people moving about the scene) and 11 (complex scenario with a large moving shadow caused by a cloud cast across the scene). All systems perform significantly worse on the second set of data, due to the more complex nature of the data. The scenes contain more people, the people being tracked appear smaller in the image, and there are heavy shadows cast by the people and the environment as well as shadows caused by moving clouds. As a result of the shadowing present, the colour modality performs very poorly, resulting in many false tracks being created as shadows from moving clouds are cast over the scene (see Figure 11, (a) - (d)). The thermal modality does not suffer from these problems, and very few false tracks are created.



(a) 50 (b) 60 (c) 70 (d) 80

(e) 50 (f) 60 (g) 70 (h) 80

(i) 50 (j) 60 (k) 70 (ℓ) 80

(m) 50 (n) 60 (o) 70 (p) 80

(q) 50 (r) 60 (s) 70 (t) 80

(u) 50 (v) 60 (w) 70 (x) 80

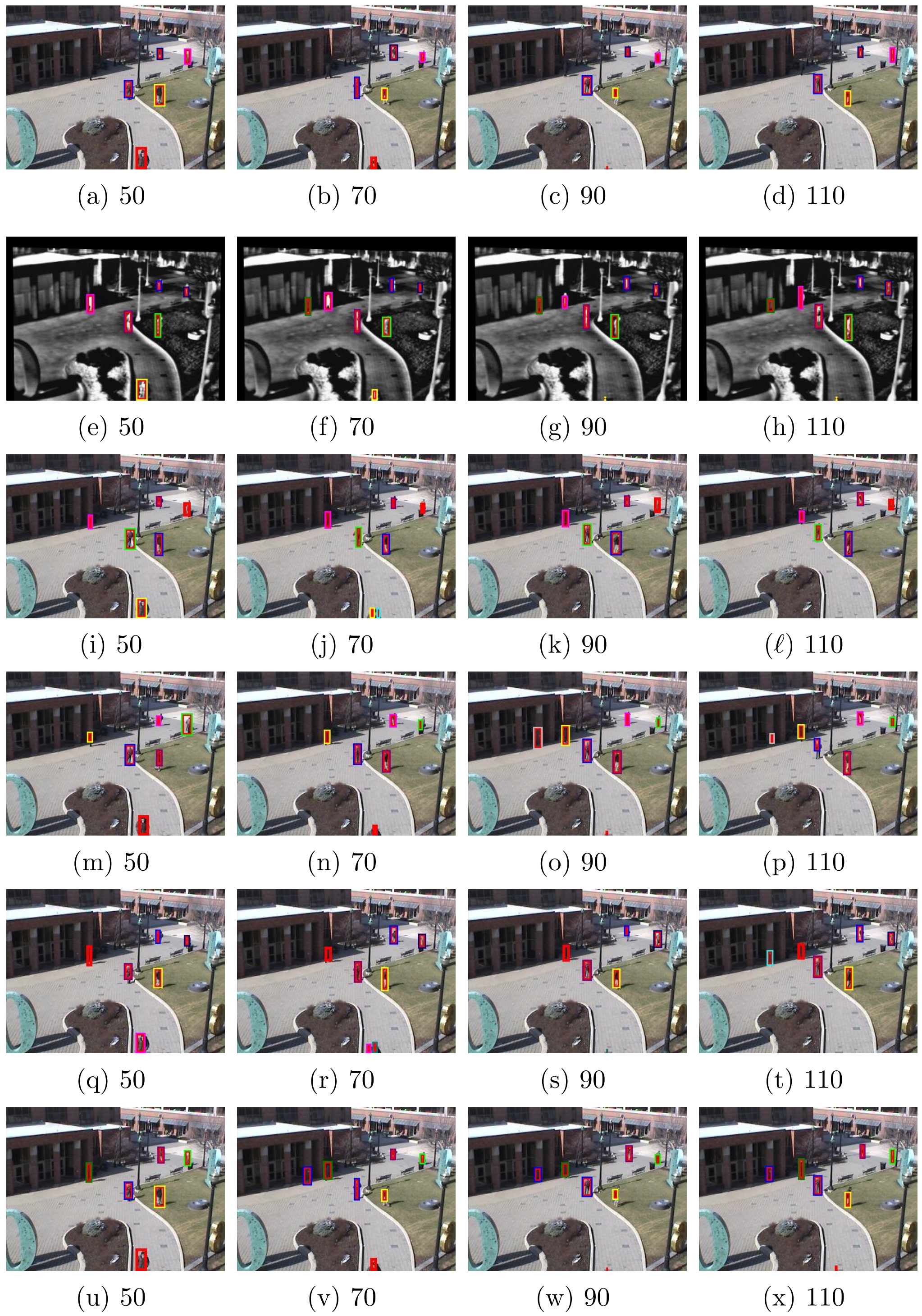
Fig. 9. Example System Results for Set 1 - Top row shows the output of tracking using colour images only; second row shows the output of tracking using the thermal images only; third row shows results of tracking using fusion scheme 1; fourth row shows results of tracking using fusion scheme 2; fifth row shows results of tracking using fusion scheme 3; sixth row shows tracking results using fusion scheme 4.

All fusion systems see some improvement over the colour modality, however all except for the third is outperformed by the thermal modality alone. The first, second and fourth fusion schemes are less effective at being able to completely ignore a modality when it is performing poorly. The first and fourth fusion schemes will always use the available information in the same manner regardless of performance. The second scheme is able to disregard an input in the event of suspected failure (the same mechanism is used by the third fusion scheme), but this will not necessarily register an error when a shadow moves gradually across the scene, and is better suited to dealing with errors caused by automatic gain control errors, or indoor situations where lights are turned on/off. The inability of Fusion schemes one, two and four to properly ignore a poorly performing mode is highlighted in Figure 11, where it can be seen that several false tracks are created as the shadow passes over the scene. This does not occur with fusion scheme three, or the thermal modality alone.

The third proposed scheme is better equipped to ignore the motion caused by shadows as it does not appear in the thermal images, and so new tracks cannot be spawned (at least some motion is required in both images to create a new track). This same mechanism also helps to deal with errors in the thermal images (see Figure 10 - in (f) a track second track is created along the building side as a result of a door being opened, however the third fusion scheme is able to avoid this).

Under appropriate conditions, all fusion schemes can offer some level of improvement over using either modality alone. Overall however, our third proposed fusion scheme (fusion after object detection) performs the best, out performing each camera on its own and the other fusion schemes. Fusions schemes one and two are directly reliant on the quality of the motion detection from the colour and thermal images. If either image contains excessive noise (sensor noise, or environmental effects such as shadowing) the whole system suffers as the fusion has been performed before any object detection processes, and so the object detection for the whole system is degraded. Fusion scheme 4 performs object detection and tracking independently on each image, and merges results. Poor performance in one modality cannot be corrected by the other modality.

Depending on the conditions of the scene, fusion schemes 1, 2 and 4 may still allow some improvement over either modality individually, however at other times it can result in reduced performance. This can possibly be overcome by modifying the early fusion schemes to determine fusion parameters dynamically, or adding additional intelligence to the multi-camera systems in fusion scheme 4 (possibly a similar system to that used in the third scheme). Fusion after the object detection overcomes this problem more effectively, as in the event that one modality produces poor results, the system can ignore this modality entirely and fall back on the second to update the system until both modalities are producing usable results.



(a) 50 (b) 70 (c) 90 (d) 110

(e) 50 (f) 70 (g) 90 (h) 110

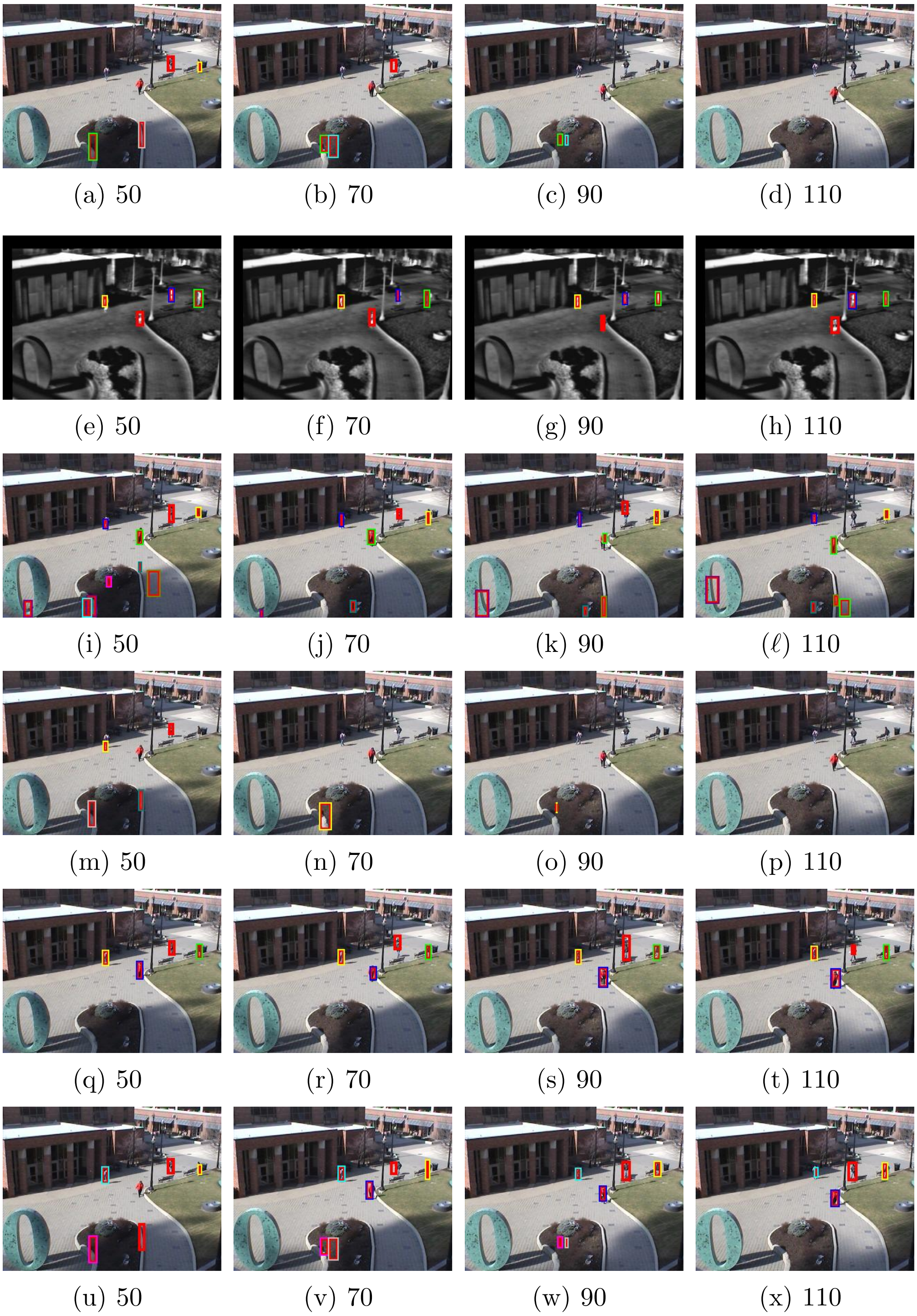
(i) 50 (j) 70 (k) 90 (ℓ) 110

(m) 50 (n) 70 (o) 90 (p) 110

(q) 50 (r) 70 (s) 90 (t) 110

(u) 50 (v) 70 (w) 90 (x) 110

Fig. 10. Example System Results for Set 2 (Example 1) - Top row shows the output of tracking using colour images only; second row shows the output of tracking using the thermal images only; third row shows results of tracking using fusion scheme 1; fourth row shows results of tracking using fusion scheme 2; fifth row shows results of tracking using fusion scheme 3; sixth row shows tracking results using fusion scheme 4.



(a) 50 (b) 70 (c) 90 (d) 110

(e) 50 (f) 70 (g) 90 (h) 110

(i) 50 (j) 70 (k) 90 (ℓ) 110

(m) 50 (n) 70 (o) 90 (p) 110

(q) 50 (r) 70 (s) 90 (t) 110

(u) 50 (v) 70 (w) 90 (x) 110

Fig. 11. Example System Results for Set 2 (Example 2) - Top row shows the output of tracking using colour images only; second row shows the output of tracking using the thermal images only; third row shows results of tracking using fusion scheme 1; fourth row shows results of tracking using fusion scheme 2; fifth row shows results of tracking using fusion scheme 3; sixth row shows tracking results using fusion scheme 4.

This approach could be further improved by applying additional intelligence to the fusion of information. The results from set two show that even when one modality (the colour modality in this case) is producing very poor results, it can still allow improvements in the detection and in the tracking over time of objects (see Table 3) due to the added colour information, which allows for better matching using appearance models and histograms, used by the condensation filter. This fusion scheme weights both inputs equally, assuming that either one is equally likely to produce valid/invalid data. The thermal modality could be weighted higher for tasks such as initial object detection to initialise tracks (so that fewer false tracks are spawned), yet the discriminating power offered by the colour modality when tracking known objects is not lost.

5.2 Abandoned Object Detection

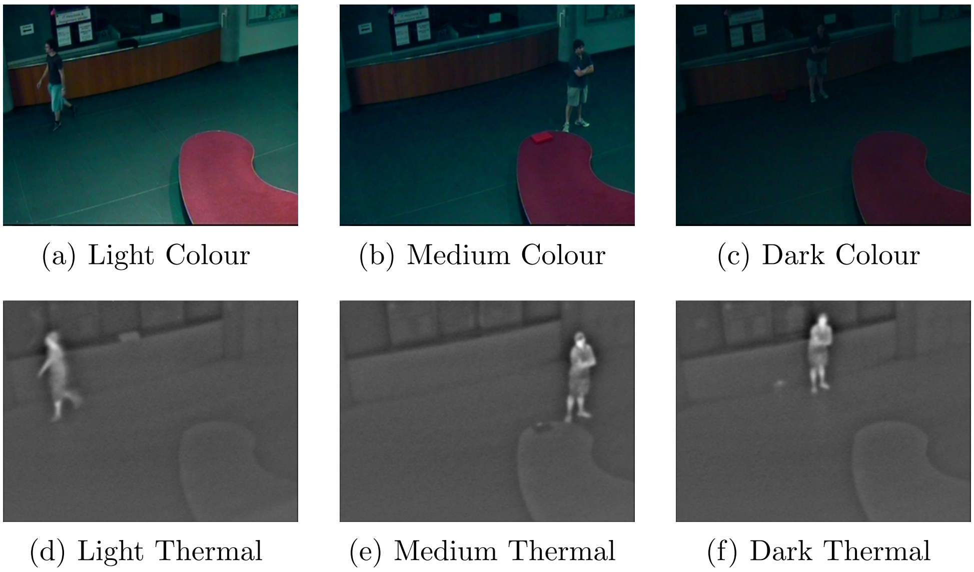
The multi-spectral abandoned object detection is evaluated using data captured in-house, as there is no publicly available multi-spectral abandoned object database. A Raytheon ControlIR 2000B IR camera, and Detection Systems V1153P colour camera were used to capture the data. Cameras were mounted on a stereo camera rig, and approximately aligned prior to capture. After capture, the images sequence are aligned and cropped.

The proposed systems are compared to systems using colour and thermal modalities individually. Performance is compared using visual inspection, as we are primarily concerned with the systems ability to detect the abandoned object and identify its owner, rather than the accuracy of the tracking system itself (this has been evaluated in Section 5.1), and this can be easily and effectively compared visually. The following items are compared in our evaluation:

(1) The ability to detect abandoned objects in terms of true positives (TP), false positives (FP) and false negatives (FN)
(2) The ability to correctly detect the owner of the object, in terms of true positives (the correct owner is identified), false positives (the incorrect owner is identified) and false negatives (no owner is identified, a valid abandoned object must have an owner)
(3) The ability to detect when an abandoned object has been collected (i.e. recognise that there is no longer an abandoned object present) in terms of true positives (the object is removed from the scene and is detected as such), false positives (the object is detect as being removed, but is actually still present) and false negatives (the object is removed from the scene, but remains detected)

For every abandoned object detection (valid or invalid), there must be a corresponding removal of the object. For a FP abandoned object, it should be removed when the object (most likely a person) causing the abandoned object moves, which results in a TP. An invalid (FP) abandoned object does not require an owner, as a FP abandoned object has no owner (any valid abandoned object does require an owner). Detecting an owner for a FP abandoned object would result in a FP for owner detection. For an abandoned object that is not detected (FN), a FN will also be recorded for owner detection and removal of object.

The database consists of abandoned object events captured in three different lighting situations (good, medium and poor light) for the colour camera, achieved by altering the shutter speed of the camera. The thermal images are captured at the same setting throughout, as thermal imaging is not effected by lighting changes. Examples of the different image conditions are shown in Figure 12. Colour and thermal images are registered prior to processing. Three sequences captured in each lighting condition are used in the evaluation. Datasets contain only one or two people, however, these people often exit and re-enter the scene. In certain scenes, the people stop and stand still for several hundred frames. It is important that false positives for abandoned object detection do not arise in these situations.


(a) Light Colour (b) Medium Colour (c) Dark Colour

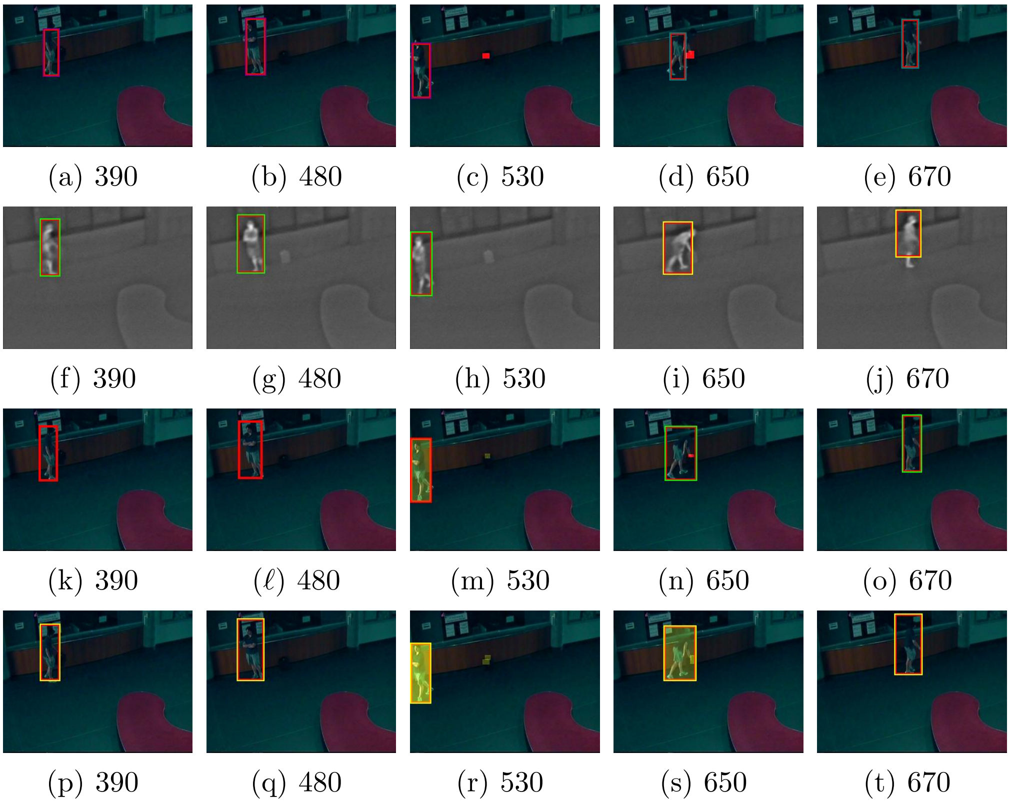
(d) Light Thermal (e) Medium Thermal (f) Dark Thermal

Fig. 12. Example Database Images

Results for the algorithms are shown in Table 4 and example output is shown in Figures 13 and 14. True positive, false positives are false negatives are shown as the number of occurrences. Within the nine sequences tested, there are 10 valid abandoned objects. Each has a single owner, and is removed from the scene. As the results in Table 4 show, the second of the proposed fusion systems (the system that uses the proposed multi-spectral abandoned object detector, the other fusion system uses early fusion) outperforms the other three systems.

Tracking System	AOD			Owner Detection			AO Removal		
	TP	FP	FN	TP	FP	FN	TP	FP	FN
Colour	7	1	3	2	0	8	8	0	3
Thermal	6	2	4	3	0	7	7	1	4
Fusion 1	7	2	3	3	0	7	9	0	3
Fusion 3	9	0	1	8	0	2	9	0	1

Table 4
Abandoned Object Detection Results



(a) 390 (b) 480 (c) 530 (d) 650 (e) 670

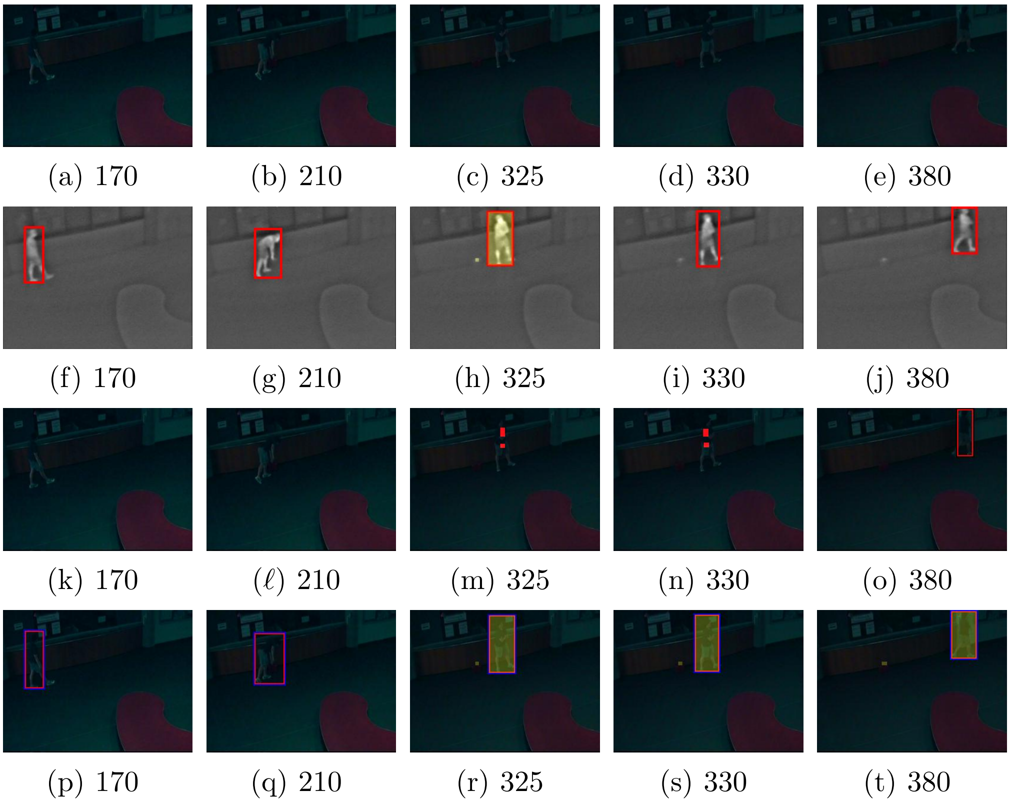
(f) 390 (g) 480 (h) 530 (i) 650 (j) 670

(k) 390 (ℓ) 480 (m) 530 (n) 650 (o) 670

(p) 390 (q) 480 (r) 530 (s) 650 (t) 670

Fig. 13. Example System Results for AOD (Medium Lighting) - Top row shows the output of tracking using colour images only; second row shows the output of tracking using the thermal images only; third row shows results of tracking using fusion scheme 1; fourth row shows results of tracking using fusion scheme 3. The colour modality is able to locate the abandoned object but not identify the owner, the thermal modality fails to detect the object. The two fusion schemes perform well.

The colour and thermal modalities performed similarly, however each was more effective in different circumstances. The colour modality failed with the darker scenes, as it was simply unable to detect motion (see Figure 14). However the thermal modality failed in situations where the temperature of the abandoned



(a) 170 (b) 210 (c) 325 (d) 330 (e) 380

(f) 170 (g) 210 (h) 325 (i) 330 (j) 380

(k) 170 (ℓ) 210 (m) 325 (n) 330 (o) 380

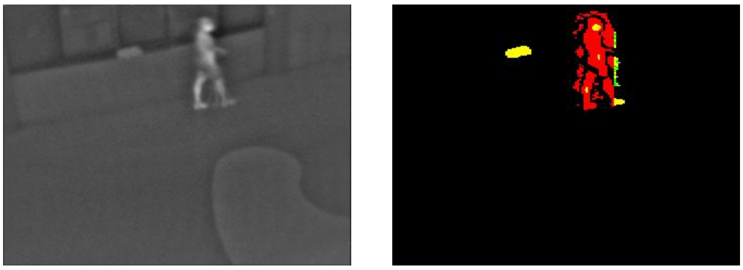
(p) 170 (q) 210 (r) 325 (s) 330 (t) 380

Fig. 14. Example System Results for AOD (Dark Lighting) - Top row shows the output of tracking using colour images only; second row shows the output of tracking using the thermal images only; third row shows results of tracking using fusion scheme 1; fourth row shows results of tracking using fusion scheme 3. The colour modality fails, while the thermal is able to detect the abandoned object and its owner briefly, but looses it before it is removed from the scene. The first fusion scheme fails to detect the abandoned object and detects to false positives on the person. The second fusion scheme (fusion scheme 3) works correctly.

object was too similar to that of the background (see Figure 13). This is in part due to the thermal halo around the moving person (see Figure 15), which restricts the thresholds that can be used for performing motion detection (it is possible that additional image processing targeted at removing this effect, could be used to improve performance).

The early fusion approach fails to improve on either modality individually, as the interlaced thermal information is not able to significantly improve performance in low light conditions (see Figure 14), and can reduce performance when the temperature of the abandoned object is similar to that of the background.

The mid fusion approach (fusion scheme three) is able to achieve a significant improvement. This approach results in nine of ten abandoned objects being detected correctly with no false positives (next best result is seven correct



(a) Input Image (b) Motion Detection

Fig. 15. Thermal Halo Example - A darker area is observed at the edge of a warmer, moving object in the scene. For the motion detection, this results in noise being detected about the edge of the moving object. If the thresholds used by the motion detection are too low, the amount of noise detected will make the tracking component of the system unusable.

detections for the colour and fusion 1 systems), and eight of the ten owners are detected correctly (next best result is three for the thermal and fusion 1 systems). This approach is able to detect the abandoned objects by observing a change in either modality, so dark scenes where colour performance is poor, or objects which are the same temperature as the scene as able to be detected (it is highly likely that an object the same colour and temperature as the background would still fail to be detected). As was seen when evaluating tracking performance (see Section 5.1), this fusion scheme also results in an improvement in tracking performance. The improvement in tracking performance results in a large increase in the correct detection of abandoned object owners.

6 Conclusions and Future Work

In this paper, we have described a multi-sensor tracking abandoned object detection systems that combine visual and thermal data to achieve better and performance than can be achieved using either mode individually. We have shown that greater improvement can be achieved by performing fusion in the later stages of the tracking process, as fusion too early can result in errors from one modality being propagated through the system. Fusing late in the process allows more control and greater flexibility over what information we choose to use or ignore. Later fusion also allows unaligned views to be used, as camera calibration information can be used to translate coordinates of detected objects between views. We have also described a novel condensation filter algorithm that allows for a more flexible, computationally efficient and robust system by allowing both the number of particles used, and types of features used to change dynamically.

Future work will focus on improving the most successful fusion scheme (fusion after object detection) proposed. The weighting of the images for different tasks such as detection and frame-to-frame tracking, as well as methods to dynamically estimate the performance of each modality will be investigated and incorporated into the system.

Acknowledgments

This project was supported by the Australian Government Department of the Prime Minister and Cabinet through the National Security Science and Technology Project Scheme.

References

[1] T. Zhao, R. Nevatia, Tracking multiple humans in complex situations, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1208–1221.

[2] I. Haritaoglu, D. Harwood, L. Davis, An appearance-based body model for multiple people tracking, in: 15th International Conference on Pattern Recognition, Vol. 4, Barcelona, Spain, 2000, pp. 184–187.

[3] L. Latecki, R. Miezianko, D. Pokrajac, Tracking motion objects in infrared videos, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2005, pp. 99–104.

[4] M. Isard, J. MacCormick, Bramble: a bayesian multiple-blob tracker, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, Vol. 2, 2001, pp. 34–41 vol.2.

[5] J. Vermaak, A. Doucet, P. Perez, Maintaining multi-modality through mixture tracking, in: Ninth IEEE International Conference on Computer Vision (ICCV'03), Vol. 2, Nice, France, 2003, pp. 1110–1116.

[6] K. Okuma, A. Taleghani, N. d. Freitas, J. Little, D. Lowe, A boosted particle filter: Multitarget detection and tracking, in: 8th European Conference on Computer Vision (ECCV), Vol. 1, Prague, Czech Republic, 2004, pp. 28–39.

[7] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: CVPR, 2001.

[8] M. Spengler, B. Schiele, Automatic detection and tracking of abandoned objects, in: VS/PETS, Nice, France, 2003.

[9] S. Guler, M. K. Farrow, Abandoned object detection in crowded places, in: IEEE International Workshop on PETS, New York, 2006, pp. 99–106.

[10] E. Auvinet, E. Grossmann, C. Rougier, M. Dahmane, J. Meunier, Left-luggage detection using homographies and simple heuristics, in: IEEE International Workshop on PETS, New York, 2006, pp. 51–58.

[11] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, R. Collins, Multi-view detection and tracking of travelers and luggage in mass transit environments, in: IEEE International Workshop on PETS, New York, 2006, pp. 67–74.

[12] C. Sacchi, C. Regazzoni, A distributed surveillance system for detection of abandoned objects in unmanned railway environments, IEEE Transactions on Vehicular Technology 49 (5) (2000) 2013–2026.

[13] E. Stringa, C. Regazzoni, Real-time video-shot detection for scene surveillance applications, IEEE Transactions on Image Processing 9 (1) (2000) 69–79.

[14] G. Foresti, L. Marcenaro, C. Regazzoni, Automatic detection and indexing of video-event shots for surveillance applications, IEEE Transactions on Multimedia 4 (4) (2002) 459–471.

[15] J. Martnez-del Rincon, J. E. Herrero-Jaraba, J. R. Gmez, C. Orrite-Uruuela, Automatic left luggage detection and tracking using multi-camera ukf, in: IEEE International Workshop on PETS, New York, 2006, pp. 59–66.

[16] C. O'Conaire, N. E. O'Connor, E. Cooke, A. F. Smeaton, Comparison of fusion methods for thermo-visual surveillance tracking, in: 9th International Conference on Information Fusion (ICIF), 2006, pp. 1–7.

[17] C. O. Conaire, N. E. O'Connor, E. Cooke, A. F. Smeaton, Multispectral object segmentation and retrieval in surveillance video, in: IEEE International Conference on Image Processing (ICIP), 2006, pp. 2381–2384.

[18] C. O. Conaire, E. Cooke, N. O'Connor, N. Murphy, A. Smearson, Background modelling in infrared and visible spectrum video for people tracking, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 3, 2005, pp. 20–20.

[19] R. S. Blum, Z. Liu, Multi-Sensor Image Fusion and Its Applications, CRC Press, Boca Raton, FL, 2006.

[20] J. Han, B. Bhanu, Fusion of color and infrared video for moving human detection, Pattern Recognition 40 (6) (2007) 1771–1784.

[21] M. Isard, A. Blake, Condensation - conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.

[22] S. Denman, V. Chandran, S. Sridharan, Person tracking using motion detection and optical flow, in: The 4rd Workshop on the Internet, Telecommunications and Signal Processing, Noosa, Australia, 2005, pp. 242–247.

[23] S. Denman, V. Chandran, S. Sridharan, An adaptive optical flow technique for person tracking systems, Elsivier Pattern Recognition Letters 28 (10) (2007) 1232–1239.

[24] S. Denman, V. Chandran, S. Sridharan, Robust multi-layer foreground segmenation for surviellance applications, in: IAPR Conference on Machine Vision Applications, Vol. 1, The University of Tokyo, Japan, 2007, pp. 496–499.

[25] A. Doucet, On sequential simulation-based methods for bayesian filtering, Technical report cued/f-infeng/tr 310, Department of Engineering, Cambridge University (1998).

[26] M. K. Pitt, N. Shephard, Filtering via simulation: Auxiliary particle filters, Journal of the American Statistical Association 94 (446) (1999) 590–599.

[27] J. Davis, V. Sharma, Ieee otcbvs ws series bench fusion-based background-subtraction using contour saliency, in: IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2005.

[28] Viper-gt, the ground truth authoring tool, http://vipertoolkit.sourceforge.net/docs/gt/.

[29] A. T. Nghiem, F. Bremond, M. T. V. Valentin, Etiseo, performance evaluation for video surveillance systems, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), London, UK, 2007, pp. 476–481.

[30] Silogic, Inria, Etiseo metrics definition (http://www-sop.inria.fr/orion/etiseo/download.htm), Tech. rep. (6th January 2006).