Improving Person Tracking Using an Inexpensive Thermal Infrared Sensor

Suren Kumar
Univ. of SUNY-Buffalo
surenkum@buffalo.edu

Tim K. Marks
Mitsubishi Electric Research Labs
tmarks@merl.com

Michael Jones
Mitsubishi Electric Research Labs
mjones@merl.com

Abstract

This paper proposes a person tracking framework using a scanning low-resolution thermal infrared (IR) sensor colocated with a wide-angle RGB camera. The low temporal and spatial resolution of the low-cost IR sensor make it unable to track moving people and prone to false detections of stationary people. Thus, IR-only tracking using only this sensor would be quite problematic. We demonstrate that despite the limited capabilities of this low-cost IR sensor, it can be used effectively to correct the errors of a real-time RGB camera-based tracker. We align the signals from the two sensors both spatially (by computing a pixel-to-pixel geometric correspondence between the two modalities) and temporally (by modeling the temporal dynamics of the scanning IR sensor), which enables multi-modal improvements based on judicious application of elementary reasoning. Our combined RGB+IR system improves upon the RGB camera-only tracking by: rejecting false positives, improving segmentation of tracked objects, and correcting false negatives (starting new tracks for people that were missed by the camera-only tracker). Since we combine RGB and thermal information at the level of RGB camera-based tracks, our method is not limited to the particular camera-based tracker that we used in our experiments. Our method could improve the results of any tracker that uses RGB camera input alone. We collect a new dataset and demonstrate the superiority of our method over RGB camera-only tracking.

1. Introduction

Person tracking is one of the fundamental problems in computer vision, and there has been extensive work on object tracking using RGB cameras. Despite much progress, human tracking remains a largely unsolved problem due to factors such as changing appearance, occlusions, motion of the camera and object, illumination variation, and background clutter [21, 22]. To deal with appearance ambiguities, a variety of methods have been proposed based on sparse representation [2], template selection and update [1], subspace-based tracking [17], and feature descriptors [16].

A fundamentally different approach to appearance ambiguities is based on using different modalities of sensing. One attractive option that has been proposed for multimodal person tracking is to use a thermal infrared (IR) camera in concert with an RGB camera[1] [10]. However, widespread adoption of thermal imaging has been hampered by the prohibitively high cost of thermal infrared cameras [18].

In this paper, we demonstrate that even a very low-cost thermal sensor can significantly improve person tracking when used in conjunction with a low-cost RGB video camera. Our thermal sensor consists of an array of 32 thermal IR receivers arranged in a vertical line, which is rotated by a motor in 94 discrete steps to produce a 140° field-of-view IR image over a time duration of 1 minute. Hence, our sensor produces a 32×94 infrared image at a rate of 1 frame per minute (0.0166 fps).

Using expensive IR cameras, tracking can be done using only thermal IR imagery [5, 19, 9]. In this paper, however, we consider what can be done with a very low-cost thermal infrared sensor, whose low resolution and extremely low frame rate preclude the possibility of tracking using IR information alone. In this paper, we will focus on person tracking in indoor scenes, in which in addition to people, there can be many heated inanimate objects such as computers, monitors and TV screens, hot drinks, and room heaters. Given the low spatial, temporal, and thermal resolution of our low-cost IR sensor, as well as variation in the temperature profile of a person due to clothing, we cannot simply use background subtraction in IR images as in [9] to determine the locations of people. Figure 1 shows an example image from our IR sensor, along with a corresponding image from the RGB camera. Only one of the four prominent blobs in the IR image corresponds to a person.

Information fusion across different modalities can be performed at various levels [3]. For example, a low-level fusion approach might combine RGB and IR information at

[1] Throughout this paper, we use the term *infrared* and the abbreviation *IR* to refer solely to thermal infrared signals, not to near-infrared (NIR) signals. We use the term *RGB camera* to refer to a video camera that operates in the visible range of the electromagnetic spectrum.

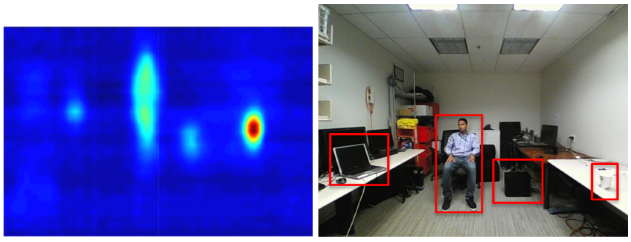


Figure 1: An image from our low-cost IR sensor (left) and a corresponding image from the RGB camera. The four blobs in the IR image correspond to (left to right): a laptop computer, a person, a CPU tower, and a cup of tea

the pixel level, before features are computed. In our case, the large gap in both spatial and temporal resolution between the RGB camera and the thermal IR sensor preclude such low-level information fusion. In a high-level fusion approach, a global decision might be reached after applying completely independent tracking algorithms in the two modalities. In this paper, we use mid-level features from the IR images to inform high-level decisions in the RGB stream.

Our system combines a real-time tracking algorithm using an RGB camera (referred to more briefly as our *RGB tracker*) with information from our IR sensor to capitalize on the strengths of both modalities, while minimizing their disadvantages. Our RGB tracker combines background modeling with template tracking. The RGB tracker is excellent at detecting moving people, but it exhibits occasional false negatives (missed detections) for stationary people and occasional false positives for inanimate objects that are moved. Due to its extremely low frame rate, our IR sensor is not useful for detecting or tracking people when they are moving about the room, and due to its low spatial resolution it cannot easily distinguish stationary people from other stationary heated objects. However, the IR sensor is extremely reliable in that it will always register stationary people as heated blobs.

By judiciously combining the low-frequency information from the thermal IR sensor with the high-level tracks from the RGB tracker, our system improves upon the RGB camera-only tracker in many situations, eliminating a variety of false positives and false negatives, and improving the region boundaries of true detections. Furthermore, the inclusion of IR information helps but does not hurt—on all of our test sequences, the incorporation of IR information does not generate any new false positives.

2. Previous Work

Here we present a brief review of tracking using three types of setup: an RGB camera alone (RGB camera-only tracking), an IR camera alone, or a combination of both IR and RGB cameras (RGB+IR).

RGB camera-only tracking We first describe three basic approaches to RGB camera-only tracking. In the first paradigm, known as *visual tracking* [21], a single object to be tracked is manually marked in the first frame of a (usually short) video sequence, then the appearance of the object and background in the first frame (along with the subsequent video frames) is used to track the object over the course of the sequence. Because visual tracking methods do not include automatic initialization of tracks, they are not complete solutions to our problem. Furthermore, visual tracking methods typically track only one object at a time, and they tend to drift over long sequences.

A second common paradigm for RGB camera-only tracking, the "tracking-by-detection" approach, provides a more complete solution for multi-person tracking. Tracking-by-detection methods rely on a person detector to find people in images, then use appearance and other cues to stitch together these detections into tracks. Such methods often use a relatively slow (not real-time) person detector and stitch together the tracks in an offline process [15].

An alternative paradigm for RGB camera-only tracking integrates detection and tracking more tightly with an online algorithm. Examples of this third paradigm include the "detect-and-track" approach of [20], which uses a background model to find candidates for tracking and couples detection and tracking in a feedback loop. In this paper, we will focus on the latter approach, as the applications we are interested in require online, real-time tracking. For a review of research related to tracking in RGB cameras, see [21, 22].

IR-only tracking Thermal IR imaging offers a tremendous advantage in differentiating people from background by virtue of temperature difference. The simplest approach, which is widely adopted, uses intensity thresholding and shape analysis to detect and track people [5]. Features traditionally used in RGB images, such as histograms of oriented gradients and other invariant features, have been adapted to IR images for person detection [14]. Recently, [9] combined background modeling in infrared with grouping analysis to perform long-term occupancy analysis.

Tracking using RGB+IR Previous approaches differ in the level at which information from the IR and RGB streams is combined. Low-level (pixel-level) combination of IR and RGB information was used by the person tracker of [13] to build a combined background model, and by [10] to improve contrast and aid in region-of-interest (ROI) segmentation by background subtraction. The system of Davis et al. [6] merges information at mid-level by first identifying ROIs in the IR domain, then obtaining contour fragments in both IR and RGB by combining gradient information. In

contrast, Zhao et al. [24] first track blobs in each modality independently, then merge the information at a high level to obtain a combined tracker.

3. Spatio-Temporal Alignment

Previous work in RGB+IR tracking uses setups in which a relatively expensive IR camera has a frame rate that is comparable to (or identical to) the frame rate of an RGB camera. Thus, previous work in this area considers only spatial alignment and does not consider temporal alignment (other than perhaps a simple matching of RGB frames to corresponding IR frames). In our setup, however, the very low-cost IR sensor is about 1800 times slower than the RGB camera, making temporal alignment of the two sensors critical. Furthermore, our IR sensor scans very slowly from side to side, capturing a single column of the IR image in the same amount of time that the camera captures multiple frames. As a result, our temporal alignment actually aligns every individual column of each IR image with corresponding columns of multiple RGB frames. (See Figure 3.)

3.1. Spatial Alignment

First it is necessary to spatially align the two cameras. For a comprehensive review of multimodal (RGB+IR) spatial image registration, we refer the interested reader to Krotosky et al. [12]. In most of the previous work on RGB+IR tracking, the outputs of the RGB and IR cameras are well approximated by a linear camera model, so they can be spatially aligned using a homography (a 3×3 linear projective transformation) between the two images [6, 24]. In our setup, both the RGB camera and IR sensor are wide-angle sensors with significant radial distortion. For this reason, a simple homography does not suffice for registering images from the two cameras. To minimize alignment problems due to depth disparities, we approximately colocated the RGB and IR cameras—that is, the two cameras were placed as close together as physically possible.

To make a calibration board for use in images from both the RGB camera and the IR sensor, we constructed a 5×3 grid of incandescent lights. (They heat up when they are left on, making them easily visible even to our low-cost, low-resolution thermal sensor.) The centers of the lights are found automatically in both the RGB and IR images by a simple blob finding algorithm constrained by the known spatial arrangement of the lights. Using the 15 corresponding points from the calibration board, we first calibrate the RGB camera and IR sensor individually and estimate their radial and tangential distortion parameters [23]. This yields nonlinear mappings, d_{rgb} and d_{ir}, that map a pixel of the raw RGB or IR image into a pixel location in the corresponding undistorted image. Next, we warp the images using the estimated distortion parameters to create IR and RGB images that are undistorted (each undistorted image obeys a linear

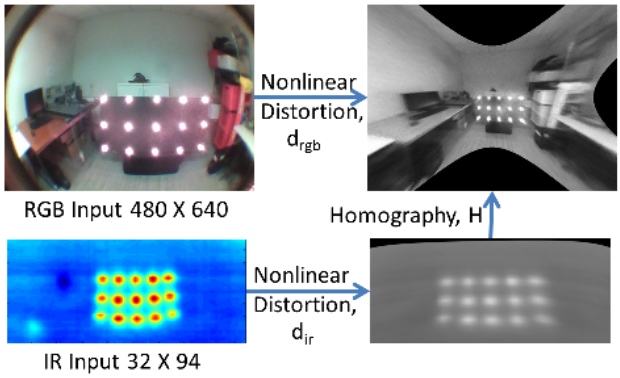


Figure 2: Spatial correspondence between images from the RGB camera and IR sensor.

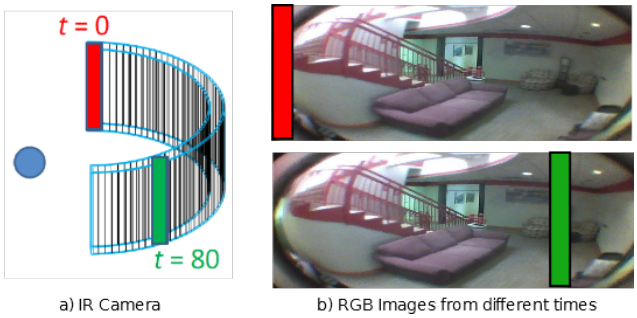


Figure 3: (a) Model of IR sensor. Over the course of a minute, the IR sensor makes a full pass from left to right, collecting 94 columns of an IR image. Two of the columns ($t = 0$ and $t = 80$) are highlighted in color. (b) Corresponding RGB images captured at $t = 0$ (top) and $t = 80$ (bottom). The IR information captured at time $t = 0$ (the leftmost column of the IR image) corresponds to the vertical stripe of the top RGB image that is highlighted in red. The IR information captured at $t = 80$ corresponds to the vertical stripe of the bottom RGB image that is highlighted in green. [Please see in color.]

camera model). The 15 correspondences between the undistorted RGB and IR images are then used to learn an infinite homography matrix, H, using Direct Linear Transformation (DLT) with isotropic scaling [11]. The registration process is illustrated in Figure 2. We represent the forward mapping from IR image to RGB image as $\mathcal{F}$ such that

$$x_{\text{rgb}} = d_{\text{rgb}}^{-1}(H d_{\text{ir}}(x_{\text{ir}})) = \mathcal{F}(x_{\text{ir}}), \tag{1}$$

where x_{rgb} is the pixel location in the RGB image corresponding to pixel location x_{ir} in the IR image.

3.2. Temporal Alignment

There has been very limited work on temporal alignment of data from IR and RGB imaging modalities, probably be-

cause the sensors in different modalities typically have similar frame rates. Conaire et al. [4] use the gen-lock input to allow two camera frame clocks to be synchronized. However, such hardware methods cannot be applied to our system because of the very low frame rate of our IR sensor. Our IR sensor uses a single column of 32 IR sensors that scan the scene in discrete steps moving from left to right to get one 140° field-of-view image, followed by a right-to-left scan to get a second 140° image. Rather than sending each column of the IR image as it is sensed, our interface to the sensor requires waiting until the end of an entire minute-long scan (a full IR image), at which time the entire IR image is transmitted. Figure 3 illustrates our model of the IR sensor. We model the dynamic motion of the IR sensor with a uniform velocity profile and use timestamps of the IR and RGB images, along with the spatial alignment described in Section 3.1, to map each column of each RGB image to a corresponding vertical stripe of the corresponding RGB frames.

This accurate spatio-temporal correspondence between the RGB camera and IR sensor is critical to our approach. For example, suppose a person walks into the scene and sits down, represented by the RGB tracker as a static track. As described in Section 5.1, when the next IR image arrives, the system verifies any static RGB track using the corresponding region in the IR image: if it corresponds to a warm blob in IR, then it is in fact a stationary person, otherwise it is a false positive. But when the IR image arrives, our system should only perform this check if the IR sensor scanned the static track's location *after* the track arrived at that location. This type of reasoning requires precise spatio-temporal correspondence.

4. RGB Tracker

Our system integrates high-level information from an RGB camera-based tracker with mid-level information (blobs) from the IR stream. Because the information from the RGB tracker is integrated at a high level (the track level), the details of the particular RGB tracker that we use are not that important. Our method for RGB+IR fusion is not restricted to the particular RGB tracker that we use—it could work with a variety of real-time, online RGB trackers. Thus in this paper, we do not give an exhaustive description of the particular RGB tracker that we use. However, in order to give a basic understanding of our RGB tracker, we briefly describe it here.

Our RGB tracker was originally developed as a standalone real-time tracking system intended for use on long video sequences of indoor living spaces. Such environments pose particular challenges that are not present in standard datasets for tracking nor for person detection, such as people in unusual poses (such as sitting down or lying on a couch), people who are stationary for a long period of time (e.g., watching TV or sleeping), people filmed from unusual perspectives (e.g., from a wide-angle camera high up in the back of a room), and lighting that is inadequate and/or changes quickly. Such video sequences cause many existing trackers and person detectors to fail. In experiments on long video sequences taken in living environments, we have found that our RGB tracker outperforms all existing tracking systems we have tested for which code is available.

We use a Gaussian-per-pixel background model to detect foreground objects in the RGB image. Detected foreground objects are tracked using a template tracker. The background model is updated at every frame, but only for pixels that are not within person tracks. Foreground detections are associated with template tracks based on overlap. Any foreground detections that do not match an existing track are treated as new detections. In order to distinguish people (which are the foreground objects that we want to track) from other foreground objects (such as new objects brought into the room, moved furniture, etc.) that we do not want to track, we use a set of visual cues. The main visual cue is motion. If an object initially moves around the room (as opposed to moving in place such as a fan or fluttering curtain), then it is assumed to be a person.

All foreground objects that are classified as people have an identity descriptor (such as a color histogram) defined for them. Matches to previous identity descriptors are another visual cue. If a newly detected foreground object is not moving around the room, then it must match a stored identity descriptor in order to be classified as a person and to continue being tracked. This visual cue handles the case in which a person walks into the room, stops moving, and remains stationary while she is occluded and then unoccluded by another person walking in front of her. Right after she is unoccluded by the person who walked in front, the stationary person is newly detected as foreground because she does not match the background model. Because her track is not moving around the room, it is required to match a stored identity descriptor in order to be classified as person. In contrast, newly detected static foreground objects that do not match a stored identity descriptor are classified as non-people and are not tracked.

These are the main visual cues that our tracker uses, although there are a few others that are of lesser importance which we do not have space to describe. Using these visual cues, our RGB tracker is able to reliably track people in indoor environments most of the time. Furthermore, using these cues helps to make our system more robust and much more computationally efficient than a state-of-the-art person detector [8].

5. Incorporating IR to Improve RGB Tracking

Although our RGB tracker works well in most cases, there are cases in which it tracks a non-person object (false positive) and cases in which the bounding box for the track

does not fit tightly around the person. Also, in certain cases our tracker may fail to track a person. For each of these failure modes, information from the low-cost IR sensor can be used to correct the problem. In general, our system tracks in real time using the RGB camera. When a new IR image becomes available (once per minute), we use warm blobs detected in the IR image to verify and improve the boundaries of static tracks and, in certain situations, to create new tracks. Because the IR sensor has such a low frame rate, it can only be applied to static tracks. The IR images cannot be used to verify or improve tracks of moving objects, since these either will not be captured by the slow IR sensor or will produce severe temporal aliasing in the IR images.

Let tr denote a particular track, and let

$$\boldsymbol{bb}_{tr}(i) = \begin{bmatrix} x_{tr}(i) & y_{tr}(i) & w_{tr}(i) & h_{tr}(i) \end{bmatrix}^{\mathrm{T}} \quad (2)$$

represent the bounding box for track tr in frame i, where $(x, y), w$, and h respectively represent the bounding box's center, width, and height. We define the *motion* of a track, tr, over the last p frames as

$$\text{motion}(tr) = \frac{1}{p} \sum_{i=f-p}^{f-1} \left\| \boldsymbol{bb}_{tr}(i) - \boldsymbol{bb}_{tr}(i+1) \right\|_1, \quad (3)$$

where $\| \cdot \|_1$ denotes the L_1 norm, and f is the index of the current frame. (In our experiments, we set p equal to one half of the ratio of the frame rates of the RGB and IR cameras.) Every track whose motion is below a threshold is considered to be a *static* track.

5.1. Non-Person Track Rejection

Each time a new IR image is obtained from the sensor, every static track that is currently present in the RGB stream is verified using the IR image, by checking that a warm blob is detected at the corresponding region in the IR image. To find warm blobs in the IR image, we simply threshold the IR image and find connected components of the set of above-threshold (warm) pixels. For each warm blob, we find the minimum enclosing bounding box, $\boldsymbol{bb}_{\text{ir}}$, in the IR image. This is mapped to the corresponding bounding box in the RGB image, $\boldsymbol{bb}_{\text{rgb}}$, by the spatial mapping, $\mathcal{F}$, described in Section 3.1. (The circumscribing rectangular bounding box of the transformed IR bounding box is used.)

To find which IR blob (if any) is associated with each static RGB track, we determine which IR blob's corresponding bounding box in the RGB frame has the greatest overlap ratio with the track's bounding box $\boldsymbol{bb}_{tr}$. For each track tr, the corresponding IR blob j^* from the set of n IR blobs in the current IR image is given by:

$$j^* = \arg \max_{j=1,..,n} \text{ov}(\boldsymbol{bb}_{tr}, \boldsymbol{bb}^j_{\text{rgb}}), \quad (4)$$

Where ov is the bounding-box overlap ratio [7] :

$$\text{ov}(\boldsymbol{bb}_1, \boldsymbol{bb}_2) = \frac{\text{area}\left(\boldsymbol{bb}_1 \bigcap \boldsymbol{bb}_2\right)}{\text{area}\left(\boldsymbol{bb}_1 \bigcup \boldsymbol{bb}_2\right)}. \quad (5)$$

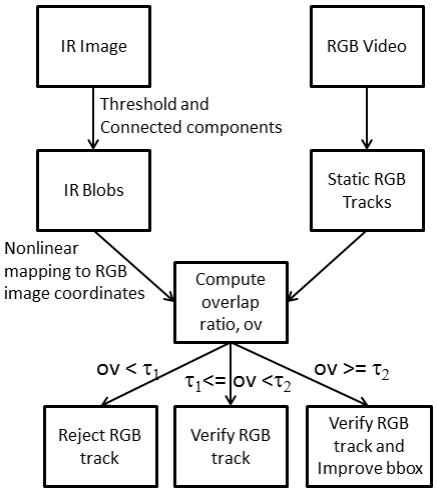


Figure 4: Flowchart summarizing how we use IR information to reject non-person tracks and to improve track bounding boxes.

If the best blob j^* has $\text{ov}\left(\boldsymbol{bb}_{tr}, \boldsymbol{bb}^{j^*}_{\text{rgb}}\right) < \tau_1$, then we reject track tr. (In our experiments, this threshold is $\tau_1 = 0.1$.)

5.2. Better Bounding Boxes

IR information can also be used to obtain better segmentation of tracked people from the background. Since the RGB tracker uses background subtraction, it can have inaccurate bounding boxes due to issues such as some foreground regions having very similar color to the local background, lighting changes, and motion blur. To improve inaccurate track bounding boxes, we replace the bounding box from the RGB tracker with the bounding box of the corrresponding IR blob if the overlap ratio (4) is greater than a threshold τ_2 (set to 0.3 in our experiments). Figure 4 shows a flowchart summary of our method for rejecting non-person tracks and improving tracks' bounding boxes.

5.3. Adding new tracks

IR information can also be used to generate new tracks. This is particularly necessary in indoor situations in which two or more people enter together such that their foreground regions overlap or touch (as in Figure 7 (a) and (b)). Since track boundaries in our RGB tracker come from background subtraction, groups of people who occlude each other when they enter the scene will be tracked as a single bounding box (since there is no appearance-based person detector to detect multiple people in a foreground blob). Such situations can commonly arise in indoor environments. An example is shown in Figure 7, in which two people enter together and sit on a couch, after which one of the people departs while the second person remains stationary on the couch. The

RGB tracker cannot infer that the remaining foreground object is actually a person, because it might be a left-behind object. (For instance, the RGB tracker cannot distinguish this situation from one in which a single person carried in a suitcase, sat down, and then departed but left his suitcase in the scene.) The remaining person is not moving, and there has been no opportunity to learn an identity descriptor for him because he has never been tracked individually.

The signature of such cases is that a track splits into two (or more) parts, and one of the parts is static and does not match any stored identity descriptors. In these cases, our RGB+IR system flags the location of the static part and stores its bounding box. When the next IR image arrives, the system checks whether there is an IR blob that intersects (overlaps with) the stored bounding box. If so, the system concludes that it must be a person and starts a new track at that location. This may seem like an unlikely scenario, but it is actually a fairly common occurrence in living environments.

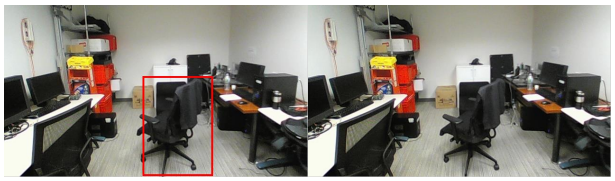
6. Experiments

We tested our method by taking video sequences using colocated IR and RGB cameras. The IR camera has a resolution of 32×94 pixels and generates 1 frame per minute. The RGB camera we use is a Genius WideCam F100 which has a resolution of 480×640 and runs at 10 frames per second with the RGB tracking algorithm running. We perform spatio-temporal calibration of our setup as explained in Section 3. With the arrival of each new IR image (once per minute), the IR information is incorporated into the tracking process as explained in Section 5.

The goal of this research is to estimate the locations of people in the image frame. We are not primarily concerned with having an accurate count of the number of people or with maintaining track identities. As explained above, our system tracks multiple people as a single entity when the foreground regions corresponding to the people intersect. For this reason, we allow a tracked bounding box to count as a match to more than one ground truth box to cover cases in which two or more overlapping people are covered by the same tracked box. To accept a tracked bounding box as true positive, we require the overlap ratio (Equation 5) between the track's bounding box and ground truth bounding box to be greater than a threshold, which we set to 0.3. We measure tracking performance in terms of two measures: detection rate (which measures the percentage of ground truth boxes tracked), and the number of false positives in each frame.

6.1. Non-Person Track Rejection

As discussed earlier, background-model-based tracking methods occasionally have false positives resulting from motion of non-person objects, such as a rolling chair. Figure 5 shows an example in which a moving object (an empty chair that is rolled into the scene and then comes to a halt) is tracked by the RGB tracker. When the next IR frame arrives, the RGB+IR system rejects that track as a non-person object (false positive) because there is no corresponding IR blob.



(a) RGB tracker (b) RGB+IR tracker

Figure 5: False positive RGB track on the left is corrected by RGB+IR tracker on the right based on the absence of an IR blob overlapping with the RGB track bounding box.

6.2. Better Bounding Boxes

Background subtraction in RGB is not always accurate due to such issues as similar colors in both foreground and background pixels, motion blur, and occlusion (which may split a single person into two foreground regions). However, as explained in Section 5.2, track boundaries can be improved by detecting IR blobs and nonlinearly transforming their bounding boxes into the space of the RGB images (see Equation 1). Figure 6 shows tracking results on a test video sequence. The graphs in (e) and (g) show detection rates per frame for the RGB tracker and RGB+IR tracker, respectively. The numbers of false positives per frame for each tracker are shown in (f) and (h). Vertical green lines in (g) and (h) indicate the instant at which each IR image arrives. In the video sequence, a person enters the scene and is tracked (see Frame 241). A rolling chair is pushed into the scene, which due to its proximity merges with the person track (as seen in Frame 2000). This creates a false positive and a missed detection for both trackers, because the overlap ratio between the track's bounding box and the ground truth is less than the threshold of 0.3. When an IR image arrives (at about frame 2100), these errors are corrected by the RGB+IR tracker when the tracked bounding box is replaced by the transformation into RGB image space of the corresponding IR blob's bounding box.

6.3. Adding new Tracks

Figure 7 shows results on another test sequence that demonstrates a scenario in which the RGB+IR tracker adds a track for a stationary person. Two people enter together in the scene and are tracked as a single entity because their boundaries touch or overlap (see Frame 750). One person gets up and walks out, and is tracked while leaving the scene (see Frame 1450). The stationary person who is left behind

is not tracked by the RGB tracker (because this case cannot be distinguished from that of a left-behind object), as shown in Frame 2000. When the next IR image arrives, the RGB+IR tracker creates a new track for the stationary person based on the rules described in Section 5.3 (see Frame 2250). Hence, the detection rate for the RGB+IR tracker goes up when the IR frame arrives, while the RGB tracker continues to fail to track the stationary person. This can be seen in the detection rate graph of Figure 7(f), in which the detection rate for the RGB+IR tracker goes up around frame 2200 while the detection rate for the RGB tracker (seen in Figure 7(e)) remains at 0. When the stationary person eventually stands up and walks away (around frame 2350), both trackers successfully track the person. The graphs of false positives per frame are not shown here, because they are almost identical for both RGB and RGB+IR trackers and contain very few false positives.

7. Conclusion

This paper presents a method for using a low-cost thermal IR sensor to improve RGB camera-based tracking. We explain how to fuse information between the IR sensor and RGB camera at the track level to remove most of the errors encountered in RGB-only tracking. We tested our framework using an RGB tracker based on integrated background subtraction and template tracking, and showed significant improvements using information from an extremely-low-frame-rate IR sensor. Our system is capable of robustly tracking people in indoor environments over long periods of time.

References

[1] N. Alt, S. Hinterstoisser, and N. Navab. Rapid selection of reliable templates for visual tracking. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1355–1362. IEEE, 2010. 1

[2] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1830–1837. IEEE, 2012. 1

[3] D. Borghys, P. Verlinde, C. Perneel, and M. Acheroy. Multi-level data fusion for the detection of targets using multi-spectral image sequences. *SPIE Optical Engineering special issue on Sensor Fusion*, 37:477–484, 1998. 1

[4] C. Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. Smearson. Background modelling in infrared and visible spectrum video for people tracking. In *CVPR Workshops*, pages 20–20, 2005. 4

[5] C. Dai, Y. Zheng, and X. Li. Pedestrian detection and tracking in infrared imagery using shape and appearance. *Computer Vision and Image Understanding*, 106(2):288–299, 2007. 1, 2

[6] J. W. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency. In *Object Tracking and Classification Beyond the Visible Spectrum, CVPR Workshops*. IEEE, 2005. 2, 3

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311, 2009. 5

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 4

[9] R. Gade, A. Jorgensen, and T. Moeslund. Long-term occupancy analysis using graph-based optimisation in thermal imagery. In *CVPR*, pages 3698–3705, 2013. 1, 2

[10] E. Goubet, J. Katz, and F. Porikli. Pedestrian tracking using thermal infrared imaging. *Infrared Technology and Applications XXXII*, pages 62062C–1, 2006. 1, 2

[11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision, Vol. 2*. Cambridge Univ Press, 2000. 3

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

[13] A. Leykin and R. Hammoud. Pedestrian tracking by fusion of thermal-visible surveillance videos. *Machine Vision and Applications*, 21(4):587–595, 2010. 2

[14] D. Olmeda, A. de la Escalera, and J. M. Armingol. Contrast invariant features for human detection in far infrared images. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 117–122. IEEE, 2012. 2

[15] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, pages 1201–1208. IEEE, 2011. 2

[16] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. In *CVPR*, volume 1, pages 728–735, 2006. 1

[17] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008. 1

[18] L. Spampinato, S. Calvari, C. Oppenheimer, and E. Boschi. Volcano surveillance using infrared cameras. *Earth-Science Reviews*, 106(1):63–91, 2011. 1

[19] J. tao Wang, D. bao Chen, H. yan Chen, and J. yu Yang. On pedestrian detection and tracking in infrared videos. *Pattern Recognition Letters*, 33(6):775 – 785, 2012. 1

[20] J. Wang, G. Bebis, and R. Miller. Robust video-based surveillance by integrating target detection with tracking. In *CVPR Workshop, 2006*, pages 137–137, 2006. 2

[21] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013. 1, 2

[22] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), 2006. 1, 2

[23] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, 2000. 3

[24] J. Zhao, S. Cheung, et al. Human segmentation by fusing visible-light and thermal imaginary. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1185–1192. IEEE, 2009. 3

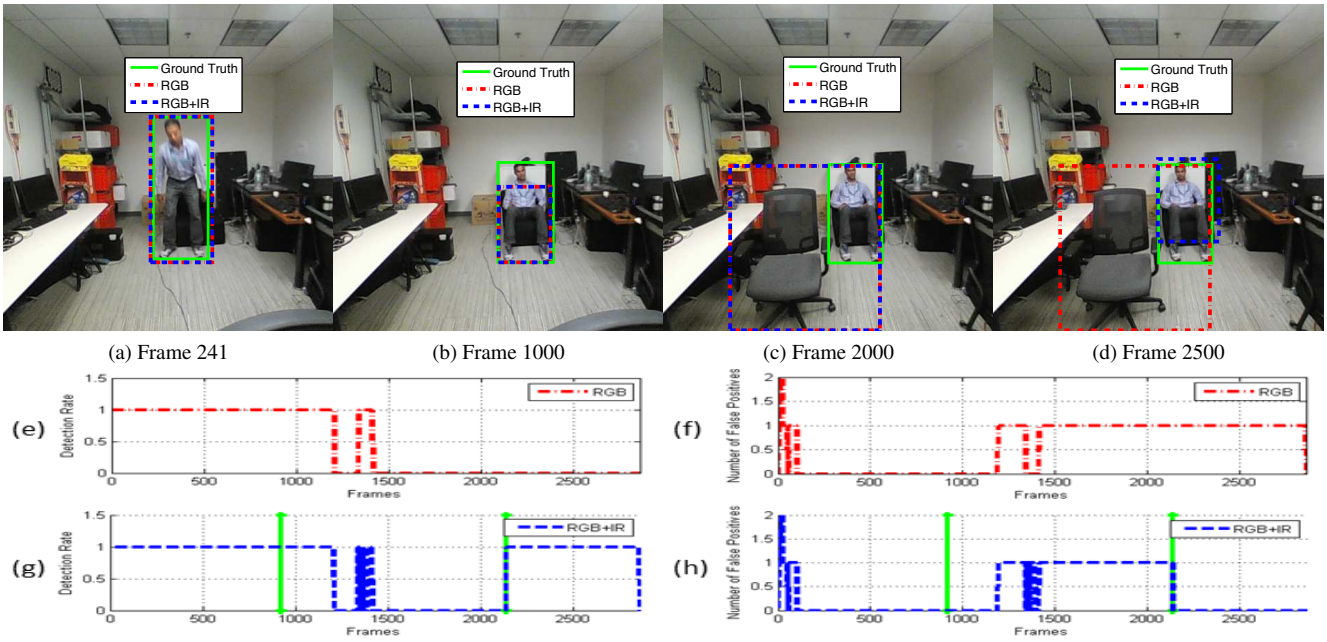


Figure 6: Images (a,b,c,d) show frames from the video with overlaid bounding boxes showing the ground truth, RGB only and RGB+IR tracks. (e,f) Detection rate and number of false positives for RGB tracking alone on Scene 2. (g,h) Same plots for RGB+IR tracking, with the arrival of each IR image indicated by a green vertical line. [Please see in color.]

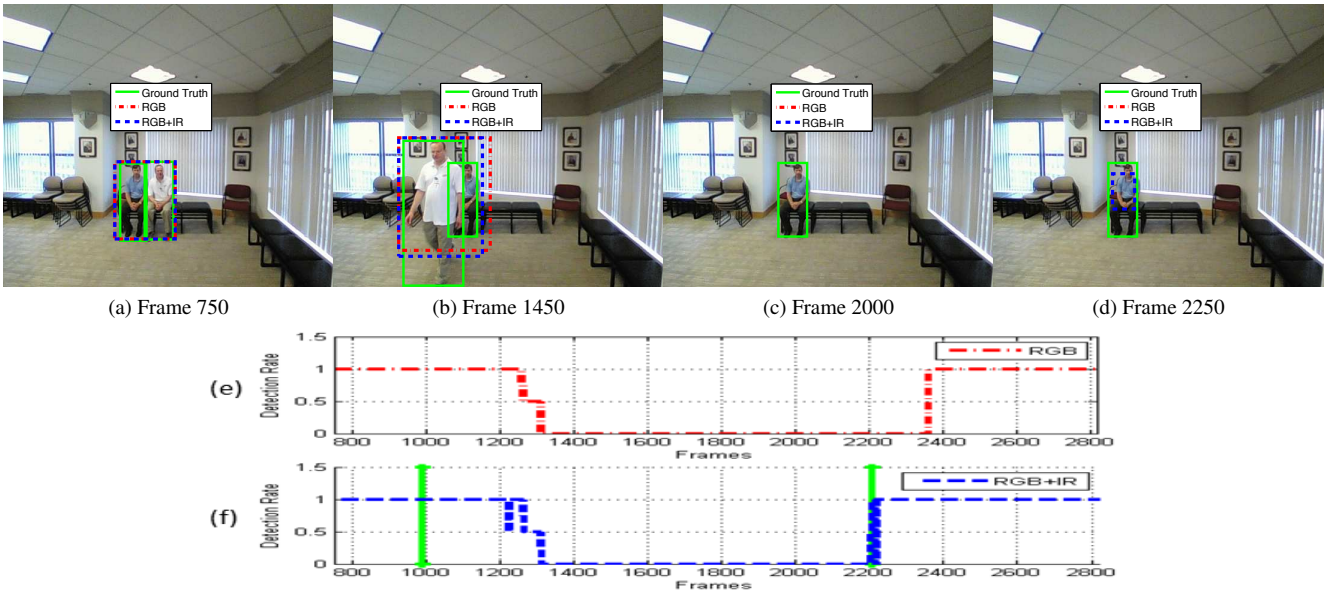


Figure 7: Images (a,b,c,d) show frames from the video with overlaid bounding boxes showing the ground truth, RGB only and RGB+IR tracks. (e) Detection rate for RGB tracking alone on Scene 3. (f) Same plot for RGB+IR tracking, with the arrival of each IR image indicated by a green vertical line. [Please see in color.]