Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 106 (2007) 162–182
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Background-subtraction using contour-based fusion of thermal and visible imagery

James W. Davis *, Vinay Sharma

Department of Computer Science and Engineering, Ohio State University, 491 Dreese Lab, 2015 Neil Avenue, Columbus, OH 43210, USA

Received 23 November 2005; accepted 15 June 2006
Available online 25 January 2007
Communicated by James Davis and Riad Hammoud

Abstract

We present a new background-subtraction technique fusing contours from thermal and visible imagery for persistent object detection in urban settings. Statistical background-subtraction in the thermal domain is used to identify the initial regions-of-interest. Color and intensity information are used within these areas to obtain the corresponding regions-of-interest in the visible domain. Within each region, input and background gradient information are combined to form a Contour Saliency Map. The binary contour fragments, obtained from corresponding Contour Saliency Maps, are then fused into a single image. An A* path-constrained search along watershed boundaries of the regions-of-interest is used to complete and close any broken segments in the fused contour image. Lastly, the contour image is flood-filled to produce silhouettes. Results of our approach are evaluated quantitatively and compared with other low- and high-level fusion techniques using manually segmented data.

Keywords: Background-subtraction; Fusion; Thermal imagery; Infrared; FLIR; Contour Saliency Map; CSM; Video surveillance and monitoring; Person detection

1. Introduction

One of the most desirable qualities of a video surveillance system is *persistence*, or the ability to be effective at all times (day and night). However a single sensor is generally not effective in all situations (e.g., a color camera at night). To attain persistence, we present a new background-subtraction technique to segment foreground objects that relies on the integration of two complementary bands of the electromagnetic spectrum, long-wave infrared (thermal) and visible light.

Thermal (FLIR) and color video cameras are both widely used for surveillance. Thermal cameras detect relative differences in the amount of thermal energy emitted/reflected from objects in the scene. These sensors are therefore independent of illumination, making them more effective than color cameras under poor lighting conditions. Color optical sensors on the other hand are oblivious to temperature differences in the scene, and are typically more effective than thermal cameras when objects are at "thermal crossover" (thermal properties of the object are similar to the surrounding environment), provided that the scene is well illuminated and the objects have color signatures different from the background.

In order to exploit the enhanced potential of using both sensors together, one needs to address the computer vision challenges that arise in both domains. While color imagery is beset by the presence of shadows, sudden illumination changes, and poor nighttime visibility, thermal imagery has its own unique challenges. The commonly used ferroelectric BST (chopper) thermal sensor yields imagery with a low signal-to-noise ratio, uncalibrated white-black polarity changes, and the "halo effect" that appears around

* Corresponding author. Fax: +1 614 292 2911.
E-mail addresses: jwdavis@cse.ohio-state.edu (J.W. Davis), sharmav@cse.ohio-state.edu (V. Sharma).

1077-3142/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.06.010

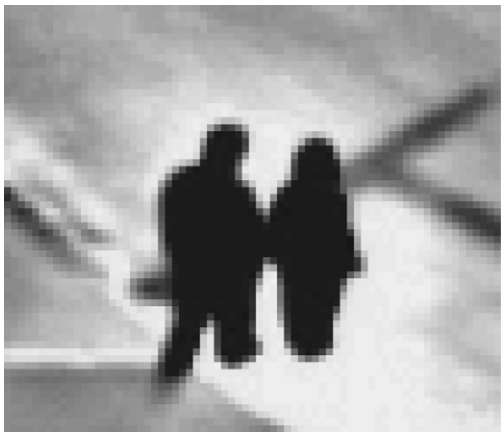


Fig. 1. Thermal image showing bright halo around dark person regions.

very hot or cold objects (see Fig. 1, showing a bright halo around dark (colder) person regions in a hot environment). The halo effect is caused due to the AC-coupling in ferroelectric focal plane arrays, that results in a droop/undershoot [29] in the response to uniformly hot and cold objects in the scene. Gain and level settings, typically used to obtain high contrast imagery with sharp object boundaries, further enhance this haloing effect and make automatic shape segmentation from thermal imagery very difficult.

These challenges of thermal imagery have been largely ignored in the past by algorithms (''hot spot'' techniques) based on the highly limiting assumption that the target object (aircraft, tank, person) is much hotter than the surrounding environment. For surveillance and other applications involving the monitoring of people, this assumption is valid only in certain conditions like cooler nighttime environments (or during Winter); it is not always true throughout the day or for different seasons of the year.

We propose an enhanced background-subtraction algorithm using both visible and thermal imagery. The approach makes use of region- and gradient-based processing to highlight contours that are the most salient within, and across, the two domains. The approach is also well-suited to handle the typical problems in both domains (e.g., shadows, thermal halos, and polarity changes). The method does not rely on any prior shape models or motion information, and therefore could be particularly useful for bootstrapping more sophisticated tracking techniques. The method is based on our previous approach [11,10] for object detection in thermal imagery.

In Fig. 2, we show a flowchart of the proposed algorithm. We start by identifying preliminary regions-of-interest (ROIs) in the two domains via standard *Background-Subtraction*. In this stage, the ROIs obtained in the thermal domain are used to localize the background-subtraction operation in the visible domain, shown by the dotted arrow in the flowchart. Next, in the *Contour Extraction* stage, we identify salient contour segments corresponding to the foreground object(s) within the ROIs of both domains by utilizing the input and background gradient information. We then *Fuse* the contours from corresponding ROIs using the combined input gradient information from both domains. In the *Silhouette Creation* stage we first close and complete the contours using an A* search algorithm constrained to a local watershed segmentation and then flood-fill the contours to create silhouettes. In the final *Post-Processing* stage we eliminate regions based on a minimum size threshold, and also use temporal filtering to remove sporadic detections. We then assign to each remaining silhouette a confidence value representative of how different it is from the background.

As shown in the figure, the entire pipe-line can be divided into three main processing levels. The *low-level* stage (Stage I) of processing deals directly with raw pixel intensities, the *mid-level* stage (Stage II) involves the extraction and manipulation of features, and the *high-level* stage (Stage III) refines the results and operates on decisions made by the lower levels. The level at which a fusion algorithm combines information from the different input sensors can play a vital role in object-detection performance. Our algorithm is a mid-level fusion technique as it fuses information at the contour level. The contour features extracted allow the algorithm to focus on high intensity

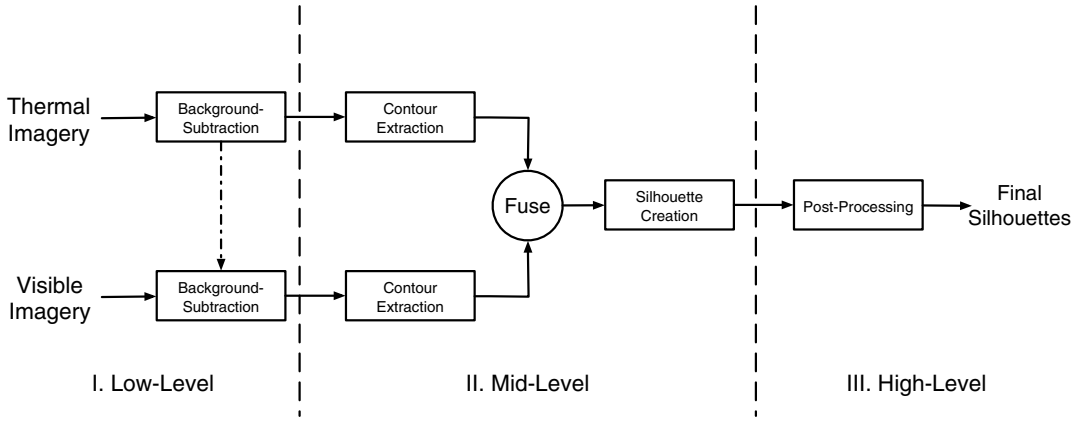


Fig. 2. Flowchart of proposed fusion algorithm.

contrasts in either domain without being affected by halos and shadows. In our method, the lower stage of the pipeline is used to focus attention on relevant image regions, and the higher stage is used only to improve the final object-detection results after fusion is performed at the mid-level.

Adopting a fusion strategy that combines information at the feature (contour) level enables us to avoid the drawbacks associated with low- and high-level fusion algorithms. Low-level algorithms combining raw intensities of pixel-regions are prone to poor results since the fusion process is unable to distinguish between image characteristics that are desirable (high contrasts) and those that are not (shadows). In high-level algorithms, the results obtained after independently processing the different input channels are usually fused using some kind of voting scheme. While this might be viable in multi-spectral imagery, such schemes do not apply well to situations when there are only *two* input channels, as is the case in the current study. Further, high-level fusion techniques are typically computationally expensive, since each input channel is processed independently until the last stage of the pipeline.

We demonstrate our approach using a single set of parameters across six challenging thermal and color video sequence pairs. The sequences, recorded from two different locations, contain large variations in the visible and thermal domains (e.g., illumination, shadows, halos, thermal gain settings, etc.). We also quantitatively analyze the results obtained by our algorithm using a set of manually segmented images. First we investigate if our algorithm indeed improves object detection performance by fusion of thermal and visible contours over using either domain independently. We then provide evidence of the shortcomings of competing low- and high-level fusion algorithms, using a comparative evaluation of object detection results.

The remainder of this paper is described as follows. We begin with a review of related work in Section 2. We then describe the main components of the proposed method. The low-level stage of our algorithm is presented in Section 3 where we describe the detection of initial regions-of-interest. In Section 4, we explain the contour extraction process and the generation of silhouettes. Section 5 then describes the last, high-level, processing stage of the pipeline, where we refine results by eliminating noisy detections. Next, in Section 6, we present experimental results. Lastly, we conclude with a summary of the research and discuss future work in Section 7.

2. Related work

Research performed in the fields of object detection and image fusion are both relevant to the current study. We first examine background-subtraction and object detection approaches proposed in both the visible and thermal domains. Related approaches in image fusion are discussed next.

2.1. Background-subtraction and object detection

In the visible domain, several object detection schemes that rely on some form of background-subtraction have been proposed. Here "foreground" regions are identified by comparison of an input image with a background model. Much research in this area has focussed on the development of efficient and robust background models. In the basic statistical approach, a distribution for each pixel (over time) is modeled as a single Gaussian [50,20], and then any new pixel not likely to belong to the distribution is detected as a foreground pixel. A Mixture of Gaussians was proposed in [45] to better model the complex background processes of each pixel. The Mixture of Gaussians approach was also examined in other work [21,36]. Other background-subtraction methods based on non-parametric statistical modeling of the pixel process have also been proposed. In [15], kernel density estimation was used to obtain the pixel intensity distributions. A variable-bandwidth kernel density estimator was proposed in [31]. Another non-parametric approach is the code-book based technique recently presented in [24]. The time series analysis of input video is another technique used to create dynamic background models. Kalman filters were used in [55], and an auto-regressive model was used in [32]. Weiner filters were employed in the three-stage (pixel/region/frame) Wallflower approach [47].

Many object detection algorithms focussing on the thermal domain directly use background-subtraction methods developed for the visible domain. The presence of halos in thermal imagery will severely impair the performance of each of the above methods as the halo artifact is typically much different than the expected background. Since the halo surrounding a person would also be detected as part of the foreground, the result would not provide an accurate localization of the person silhouette, when ironically the person shape is most easily distinguishable to human observers *because* of the halo (halos are at opposite polarity to the object [13]). Some of these methods (e.g., [20,19]) have been tested with thermal imagery, but the limited nature of the examples examined does not provide a comprehensive evaluation.

The unifying assumption in most other person-detection algorithms aimed at the thermal domain is the belief that humans are warmer than their surroundings, and hence appear brighter, as "hot-spots". In [23] and [2], a thresholded thermal image forms the first stage of processing after which methods for pose estimation and gait analysis are explored. In [34], a simple intensity threshold is employed and followed by a probabilistic template. A similar approach using Support Vector Machines is reported in [51]. The use of the strong hot-spot assumption can also be found in other work related to object detection and tracking in thermal imagery [3,7,52]. The underlying hot-spot assumption will be violated in imagery recorded at different environmental temperatures and in most urban environments.

2.2. Fusion

Image fusion techniques have had a long history in computer-vision and visualization. We categorize related work into three types, based on the processing level (low, mid, high) at which fusion is performed.

Traditionally, *low-level* techniques have been used to combine information from co-registered multi-sensor imagery. Improving upon simple techniques such as pixel averaging, multi-resolution schemes similar to the pyramid-based approaches of [46,35,6] were proposed. More recently, wavelet analysis has emerged as the method of choice in most multi-resolution frameworks [28,39]. Examples of other low-level techniques include the biologically motivated model based on human opponent color processing proposed in [17]. A PCA-based technique measuring pixel variances in local neighborhoods is used in [8]. Pixel-level combinations of spatial interest images using Boolean and fuzzy-logic operators are proposed in [14], and a neural networks model for pixel-level classification is used in [25].

Mid-level fusion techniques have mostly relied on first and second order gradient information. Some of these techniques include directly combining gradients [40], determining gradients in high dimensions [44], and analyzing gradients at multiple resolutions [43,37]. Other features, such as the texture arrays [5], have also been employed. Model-based alternatives to feature-level fusion have also been proposed such as the adaptive model matching approach of [16], and the model-theory approach of [53]. Other mid-level fusion techniques such as the region-based methods of [27,54,38] make use of low-level interactions of the input domains.

High-level fusion techniques generally make use of Boolean operators or other heuristic scores (maximum vote, weighted voting, m-of-n votes) [9,48] to combine results obtained from independently processing the input channels. Other "soft" decision techniques include Bayesian inference [1,22] and the Dempster–Shafer method [4,30].

Most of these fusion techniques aim at enhancing the information content of the scene, to ease and improve human interpretation (visual analysis). However, the method we propose is designed specifically to enhance the capabilities of an automatic vision-based detection system. Some techniques such as [14,16,5], proposed for Automatic Target Recognition systems, have also been evaluated in terms of object detection performance. These techniques however are not generally applicable to the detection of non-rigid person shapes, and other large, multi-modal objects common in the urban environments considered in this work. Other techniques, such as [17], have been shown to improve recognition performance when used as inputs to separate target recognition modules. In contrast, our approach tightly integrates the fusion of information with the process of object detection, thus resulting in a *single* pipeline that exploits different sensors for improving object detection.

3. Stage I: low level processing

In this initial stage of processing, our algorithm starts by registering the two input streams. Then, using standard background-subtraction, our algorithm identifies initial regions-of-interest in the thermal domain. These are then used to cue the selection of corresponding regions from the visible domain.

3.1. Image registration

Our algorithm requires the two input image streams to be registered. For a particular camera location, we first manually select a set of corresponding feature points from a pair of thermal and visible images. Using these points we compute the homography matrix [18] to register each thermal image of the sequence with the corresponding visible image. In Fig. 3(a) and (b), we show a pair of visible and thermal images. Fig. 3(c) shows the thermal image after registration. In Fig. 3(d), we show the result of a pixel-wise max operator used to combine the images *before* registration. Note the large mis-alignment in the two views. The result *after* registration is shown in Fig. 3(e).

3.2. Initial region detection

We begin by identifying regions-of-interest (ROIs) in both domains (thermal and visible). The background in the thermal domain tends to be more stable over time, as it changes more slowly with environmental variations. Standard background-subtraction in the thermal domain is thus more reliable and generally produces regions that encompass the entire foreground object and the surrounding halo. Therefore we first use background-subtraction to obtain the ROIs in the thermal domain, and extract the corresponding ROIs from the visible domain. We use a single Gaussian at each pixel to model the background. Other statistical approaches to model the background, such as Mixture of Gaussians [45] or code-book techniques [24], could also be used, but will not be sufficient to address the halo artifact in thermal imagery and the presence of object shadows in visible imagery. As discussed in [13], halos and object shadows co-occur with foreground objects. Hence, just as foreground objects appear distinct from the background, so too would the halos and shadows caused by them.

To bootstrap the construction of proper mean/variance background models from images containing foreground objects, we first capture N images in both the thermal and visible domains. We begin by computing a median image (I_{med}) from the N thermal and visible intensity images. The statistical background model for each pixel (in thermal or visible intensity) is created by computing *weighted* means and variances of the N sampled values

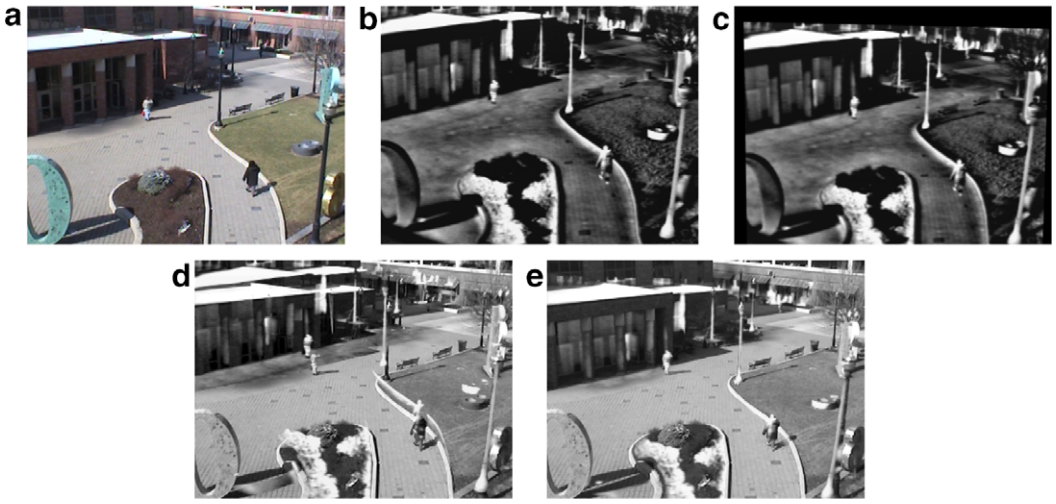


Fig. 3. Image registration. (a) Visible image. (b) Original thermal image. (c) Registered thermal image. (d) Pixel-wise max of images (a) and (b). (e) Pixel-wise max of images (a) and (c).

$$\mu(x,y) = \frac{\sum_{i=1}^{N} w_i(x,y) \cdot I_i(x,y)}{\sum_{i=1}^{N} w_i(x,y)} \tag{1}$$

$$\sigma^2(x,y) = \frac{\sum_{i=1}^{N} w_i(x,y) \cdot (I_i(x,y) - \mu(x,y))^2}{\frac{N-1}{N} \cdot \sum_{i=1}^{N} w_i(x,y)} \tag{2}$$

where the weights $w_i(x,y)$ for a pixel location are used to minimize the effect of outliers (values far from the median $I_{\text{med}}(x,y)$). The weights are computed from a Gaussian distribution centered at $I_{\text{med}}(x,y)$

$$w_i(x,y) = \exp\left(\frac{(I_i(x,y) - I_{\text{med}}(x,y))^2}{-2\widehat{\sigma}^2}\right) \tag{3}$$

The farther $I_i(x,y)$ is from $I_{\text{med}}(x,y)$, the smaller its contribution. In our experiments, we used a standard deviation $\widehat{\sigma} = 5$ (a pixel-wise $\widehat{\sigma}$, learned from the N frames could also be used). Using these weights to compute the statistical background model enables us to obtain strong background models without requiring training images to be completely devoid of foreground activity. For longer sequences, the background model can be updated using schemes, as in [45], with

$$\begin{aligned}\mu_t(x,y) &= (1-\rho) \cdot \mu_{t-1}(x,y) + \rho \cdot I_t(x,y) \\ \sigma_t^2(x,y) &= (1-\rho) \cdot \sigma_{t-1}^2 + \rho \cdot (I_t(x,y) - \mu_t(x,y))^{\mathrm{T}} \\ &\quad \times (I_t(x,y) - \mu_t(x,y))\end{aligned} \tag{4}$$

where the subscript t denotes time and ρ is the update factor (typically $\rho \ll 1$).

Having computed the statistical background model for the thermal domain (using Eqs. (1) and (2)), we obtain the foreground pixels, D^{T}, for an input thermal image using the squared Mahalanobis distance

$$D^{\mathrm{T}}(x,y) = \begin{cases} 1 & \frac{(I(x,y)-\mu(x,y))^2}{\sigma(x,y)^2} > Z^2 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Since further processing in either domain is limited to the foreground pixels chosen at this stage, a liberal threshold (Z) that yields all of the object regions (and portions of the background around the object) is preferred over a more aggressive setting (perhaps deleting portions of the foreground object). Several different thresholds and their effect on the overall performance are examined in our experiments (Section 6). To extract the thermal ROIs from the selected foreground pixels, we apply a 5×5 dilation operator to the background-subtracted image D^{T} and employ a connected components algorithm. Any region with a size less than approximately 40 pixels is discarded (for a 320×240 image).

Ideally, we could simply use D^{T} to represent the ROIs in the visible domain, D^{V}, as well. However, the image regions in the visible domain corresponding to D^{T} may contain unwanted artifacts such as shadows, or may be void of intensity or color information differentiating them from the background. Thus, for the visible image of the input image pair, we perform an additional color/intensity background-subtraction *within* image regions identified using D^{T}. While any background-subtraction scheme that minimizes the effect of shadows can be used, we employ a simple technique that exploits the fact that a shadow lowers surface luminosity without changing chrominance characteristics.

For each visible image region corresponding to a region in D^{T}, the intensity component is used to identify pixels (D_{Int}) that are statistically brighter than the background, and the normalized RGB components are used to detect pixels (D_{Col}) different in color from the background. For the intensity component, a mean/variance model is computed using Eqs. (1) and (2), while the mean/covariance model of the normalized color-space is computed directly from the initial set of N visible images (without the weights in Eq. 3) and can be updated over time using Eq. 4. The visible domain ROIs are then obtained by a pixel-wise union of D_{Int} and D_{Col} followed by a 5×5 dilation operator, similar to the thermal domain.

For the pair of input images in Fig. 4(a) and (b), we show the computed D^{T} and D^{V} in Fig. 4(c) and (d), respectively. The squared Mahalanobis distance thresholds for the thermal, luminance, and color channels were set at 10, 15, and 4, respectively for this example. Using D^{T} to localize the identification of ROIs in the visible domain enables us to use thresholds that maximize detection of object pixels within specific regions without incurring any penalty for false detections that occur elsewhere in the image (thus the thresholds need not be precise). The result of applying background-subtraction to the *entire* visible image, with the same liberal threshold as used for Fig. 4(d), is shown in Fig. 4(e). In spite of using the same thresholds, intensity fluctuations and other image/sensor characteristics result in a large number of spurious detections. However, using the thermal ROI (Fig. 4(d)) as a mask permits a reasonable result in the visible image without the large number of false positives. If the thresholds were changed to reduce the spurious detections in Fig. 4(e), more person regions will be lost, which would be detrimental. Fine tuning the background-subtraction method could be used to reduce the false positives, but this would have to be done on a case-by-case basis. Using the thermal ROIs as a mask enables us to set a liberal and generalized threshold that ensures detections in the visible domain correspond mostly to the desired person regions. Obviously, more detailed background-subtraction approaches (including shadow removal) could be used, but as we will show, the quality of detection obtained by this overlay method is adequate. Having obtained D^{V}, the subsequent stages of our algorithm make use of only the intensity components of both the thermal and visible domains.

4. Stage II: mid-level processing

In this stage of the pipeline, our algorithm relies on (input and background) *gradient* information within the selected ROIs. First these gradients are manipulated to highlight salient object boundaries in the thermal and visible domains. Binary contour features are then extracted from this representation in either domain and combined into single fused contour image. These contours are then closed and completed to form silhouette regions.

4.1. Contour extraction

We first examine each ROI in the thermal and visible domains individually in an attempt to extract gradient information corresponding only to the foreground object. For each ROI, we form a *Contour Saliency Map* (CSM) [10], where the value of each pixel in the CSM represents the confidence/belief of that pixel belonging to the boundary of a foreground object.

A CSM is formed by finding the pixel-wise minimum of the normalized input gradient magnitudes and the normalized input-background gradient-difference magnitudes within the ROI

$$\mathrm{CSM} = \min\left(\frac{\|\langle I_x, I_y\rangle\|}{\mathrm{Max_I}}, \frac{\|\langle (I_x - \mathrm{BG}_x), (I_y - \mathrm{BG}_y)\rangle\|}{\mathrm{Max_{I-BG}}}\right) \quad (6)$$

where I_x and I_y are input gradients, BG_x and BG_y are background gradients, and the normalization factors, $\mathrm{Max_I}$ and $\mathrm{Max_{I-BG}}$, are the respective maximum magnitudes of the input gradients and the input-background gradient-differences in the ROI. The range of pixel values in the CSM is $[0,1]$, with larger values indicating stronger confidence that a pixel belongs to the foreground object boundary.

The motivations for the formulation of the CSM are that it suppresses (1) large non-object input gradient magnitudes (as they have small input-background gradient-difference magnitudes), and (2) large non-object input-background gradient-difference magnitudes (typically from thermal halos or diffuse visible shadows). Thus, the CSM preserves the input gradients that are both

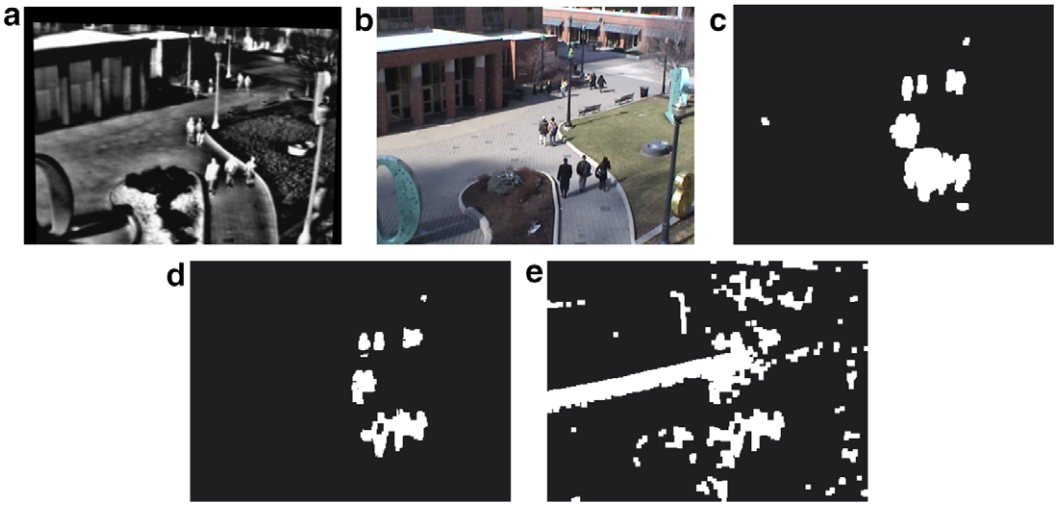


Fig. 4. Region detection. (a) Input thermal image. (b) Input visible image. (c) D^{T}. (d) D^{V}. (e) Background-subtraction results in the visible domain using same thresholds as in (d).

strong *and* significantly different from the background. The approach is equally applicable to both thermal and visible imagery. We compute the CSM for all ROIs in both the thermal and visible (intensity) domains.

We show the CSM construction for a thermal and visible ROI in the top and bottom rows Fig. 5, respectively. The gradients were calculated using 7×7 Gaussian derivative masks.

Fig. 6. CSM thinning. (a) CSM. (b) Non-maximum suppression of input gradient magnitudes. (c) tCSM.

4.1.1. Thinning

Our next step is to produce a thinned (1-pixel thick contours) representation of the CSM, which we call the tCSM. As the CSM does not represent a true gradient image, standard non-maximum suppression methods that look for local peaks along gradient directions (as used in the Canny edge detector) cannot be directly applied. However, by the composite nature of the CSM, maxima in the CSM must always co-occur with maxima in the input gradients. Therefore we can use the non-maximum suppression result of the *input* gradients as a thinning mask for the CSM. In Fig. 6, we show a CSM, the non-maximum suppression thinning mask (derived from the input gradients in Fig. 5(b)), and the final tCSM computed from the multiplication of the CSM with the thinning mask.

4.1.2. Thresholding

After thinning, we threshold the tCSM to select the most salient contour segments. We use the competitive clustering technique of [13] in both the thermal and the visible domains. Though the approach was originally motivated by object properties in the thermal domain, it is not exclusive to thermal imagery. The technique first clusters tCSM pixels based on saliency and then discards the lowest cluster. To ensure that the lowest cluster contains background contour fragments, it was observed in [13] that it is best to use 2 clusters when the object regions are unimodal and 3 when they are multimodal. This observation holds irrespective of the nature (thermal or visible) of the imagery.

Instead of trying to estimate the modality of the input, every tCSM is clustered (using *K*-means) twice, into 2 and 3 saliency groups corresponding to the unimodal and multimodal cases, and thresholded by setting all pixels in the lowest cluster to 0 (the remaining pixels are set to 1). The cluster centers in the *K*-means algorithm are initialized to the minimum and maximum tCSM values in the two cluster case, and to the minimum, median, and maximum tCSM values in the three cluster case. The optimal binary image is then chosen from the two thresholded tCSMs, B_2 and B_3. To rank the two binary images we form a quality measurement Q using the *average contour length* (ACL) and *coverage* (C). The ACL is computed by averaging the lengths of the individual contours obtained from a region-growing procedure applied to the thresholded tCSM. We use the average distance of the perimeter pixels of the *ROI* to the closest pixel in the thresholded tCSM as an *inverse* measure of C (a lower C thus indicates better coverage). The hypothesis is that an optimally thresholded tCSM should contain contours of relatively high average length that also "cover" the ROI sufficiently well.

Thus the quality of a thresholded image B_i is evaluated using

$$Q(B_i) = (1-\alpha)\cdot\left(\frac{\mathrm{ACL}(B_i)}{\max(\mathrm{ACL}(B_2),\mathrm{ACL}(B_3))}\right) + \alpha\cdot\left(1-\frac{C(B_i)}{\max(C(B_2),C(B_3))}\right)$$

The binary image (B_2, B_3) that maximizes Q is chosen as the best thresholded result, which we represent as $\mathrm{tCSM_b}$, the subscript b denoting that the tCSM is binary.

Essentially, Q is a weighted sum of the normalized ACL and coverage values. The weighting factor α determines the

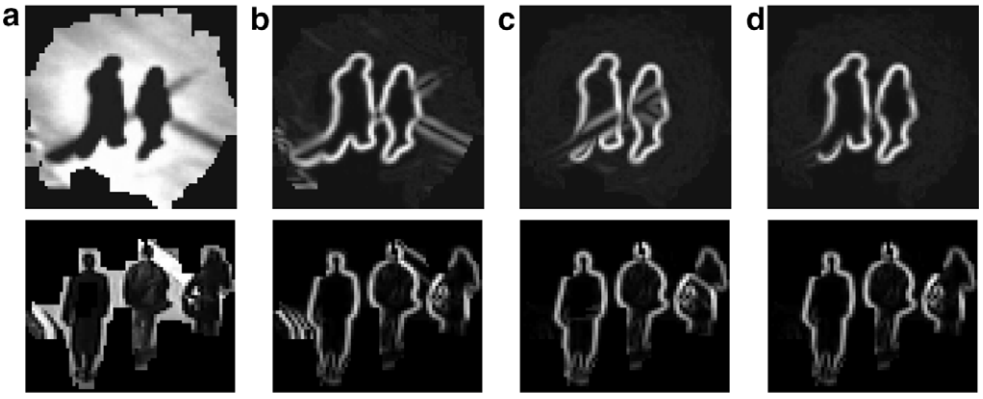


Fig. 5. Contour saliency in the thermal (top row) and visible (bottom row) domains. (a) ROI. (b) Input gradient magnitudes. (c) Input-background gradient-difference magnitudes. (d) CSM.

influence of each of the factors on Q. Empirically, we found that if the ACL of one of the images is less than half of the other, then there is little need to rely on C. On the other hand, if the two ACLs are quite similar, then C should be the most influential factor. In other words, the weight α should be a function of the ratio of the two ACLs

$$r = \frac{\min(\mathrm{ACL}(B_2), \mathrm{ACL}(B_3))}{\max(\mathrm{ACL}(B_2), \mathrm{ACL}(B_3))} \tag{7}$$

and, when $r > 0.5$, α should be $\sim$1, and when $r < 0.5$, α should be $\sim$0. We therefore express α non-linearly as a sigmoid function centered at 0.5 given by

$$\alpha = \frac{1}{1 + \mathrm{e}^{-\beta \cdot (r - 0.5)}} \tag{8}$$

where the parameter β controls the sharpness of the non-linearity (we use $\beta = 10$).

In Fig. 7(a)–(c), we show a thermal ROI with *unimodal* person pixels and the competing binary images, B_2 and B_3, respectively. The resulting quality values are $Q(B_2) = 0.993$ and $Q(B_3) = 0.104$. Thus, as expected due to the unimodal nature of the person pixels, B_2 was selected as the correct thresholded image. In this example, the large difference in the ACL ($r = 0.154$) of the two images resulted in it being the dominating factor in the quality evaluation. In Fig. 7(d)–(f), we show a thermal ROI with *multimodal* person pixels and its binary images, B_2 and B_3. The resulting quality values were $Q(B_2) = 0.103$ and $Q(B_3) = 0.255$, and as expected, B_3 was correctly selected. The dominant quality factor here was the coverage, since the ACLs were similar ($r = 0.893$).

Fig. 8 shows examples of the thresholding technique for the visible domain. A unimodal ROI is shown in Fig. 8(a), with the competing binary images B_2 and B_3 shown in Fig. 8(b) and (c), respectively. As shown, the competing binary images are similar, and their quality values are $Q(B_2) = 0.184$ and $Q(B_3) = 0.395$. Here, unlike in the thermal domain (Fig. 7(a)–(c)), we find that both candidates had almost equal ACLs ($r = 0.752$), thus making C the dominating factor. This can be attributed to the absence of the halo artifact in visible imagery, which ensures that, when objects appear unimodal, the tCSM values are tightly clustered around 0 (background) and 1 (object gradients). Thus, unlike the thermal domain, the middle cluster in B_3 does not contain noise fragments, and in fact contributes to the better quality of B_3. However, the presence of diffused shadows could simulate conditions similar to halos in the thermal domain, and the middle cluster would then contain unwanted contour fragments (affected by the shadow). Since we employ a shadow-removal stage prior to the formation of the tCSM, this scenario seldom occurs, and both B_2 and B_3 are generally good candidates for the final binary result. In Fig. 8(d), we show a visible ROI with *multimodal* person pixels. The resulting quality values of its binary images, shown in Fig. 8(e) and (f) are $Q(B_2) = 0.316$ and $Q(B_3) = 0.674$, respectively. Similar to the corresponding case in the thermal domain, the ACLs were comparable ($r = 0.714$), and the dominant quality factor here again was C.

4.2. Contour fusion

We now have binary contour fragments corresponding to the same image region in both the thermal and the visible domains. Within their respective domains, these contours lie along pixels with the most salient object gradients. Based on the complementary nature of the sensors, we anticipate that these salient contours, when combined, would provide a better, less broken, delineation of the object boundary. Our next task is thus to combine the contours into a single fused image from which we will form silhouettes.

Since the tCSMs of either domain *only* contain contours that reliably represent salient object boundaries, the fused result can contain *all* contour features from both domains. We thus combine information from the two sensors by performing a simple union of their individual contributions using

$$\mathrm{tCSM_b} = \mathrm{tCSM_b^T} \cup \mathrm{tCSM_b^V} \tag{9}$$

While this provides a simple way to benefit from the complementary nature of the sensors, sensor redundancy can sometimes be problematic. Contour fragments from the thermal and visible domain belonging to the same edge may not always perfectly coincide with each other (due to imperfect registration, differences in the sensors, etc.). In order to complete and close these contour fragments for the creation of silhouettes, we require that the contour fragments be only 1-pixel thick. Hence, if two contour fragments lie adjacent to each other (instead of over each other) in the $\mathrm{tCSM_b}$, we would like to preserve only the contour corresponding to the stronger gradient. The final $\mathrm{tCSM_b}$ therefore needs to be further "aligned" such that only those contour fragments that correspond to gradient maxima across both domains are preserved.

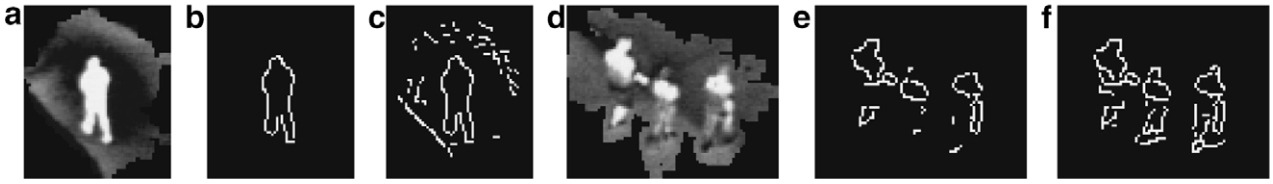


Fig. 7. Contour selection in thermal domain. (a) Unimodal ROI. (b) B_2 (selected). (c) B_3. (d) Multimodal ROI. (e) B_2. (f) B_3 (selected).

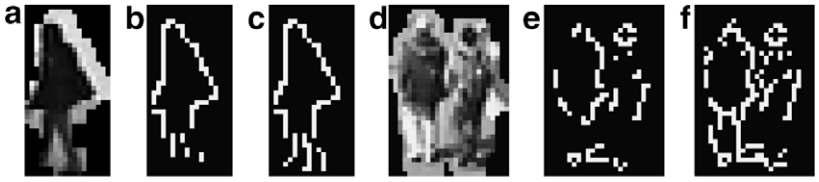


Fig. 8. Contour selection in visible domain. (a) Unimodal ROI. (b) B_2. (c) B_3 (selected). (d) Multimodal ROI. (e) B_2. (f) B_3 (selected).

To achieve this, we first create a combined input gradient map from the foreground gradients of each domain. Gradient direction and magnitude information for a pixel in $\mathrm{tCSM_b}$ is selected from either the thermal or the visible domain depending on it being present in the $\mathrm{tCSM_b^T}$ or $\mathrm{tCSM_b^V}$ (if present in both, the gradient information at that pixel can be taken from either domain). Since we now have an orientation and magnitude at every contour pixel in the $\mathrm{tCSM_b}$, we apply a local non-maximum suppression algorithm to perform a second thinning to better align the $\mathrm{tCSM_b}$. This results in a set of contours that are the most salient in the individual domains as well as *across* the domains. In Fig. 9(a) and (b), we show corresponding ROIs in the thermal and visible domains. Fig. 9(c) and (d) shows the $\mathrm{tCSM_b}$ before and after alignment, respectively. In Fig. 9(e), we show the final contours obtained after alignment (Fig. 9(d)) overlaid on the result of the raw union of contours (Fig. 9(c)) from the thermal and visible domains. While in the top row of Fig. 9(c) the contour image appears complete, in most cases the union of contours from the two sensors leaves several gaps in the contour image (as shown in the corresponding figure in the bottom row). The contours in Fig. 9(d) appear more "broken" than in Fig. 9(c), but since they are 1-pixel thick, they can now be effectively completed and closed.

4.3. Contour completion and closing

While contour information from the two channels are often complementary, the contour fragments in the combined $\mathrm{tCSM_b}$ are still mostly broken (see Fig. 9(d)) and need to be *completed* (i.e., the contours have no gaps) and *closed* (i.e., the contour figure is equivalent to the closure of its interior) before we can apply the flood-fill operation to create silhouettes. To achieve this, we use the two-part method originally proposed in [11,10]. The algorithm first attempts to connect any gaps in the figure using an A* search algorithm to grow out from each gap endpoint towards another contour pixel. Next, all contours in the figure are closed.

To limit the search space and constrain the solution to have meaningful path completions/closings, the method makes use of the Watershed Transform (WT) [49]. When the WT is applied to a gradient magnitude image, the resulting watershed lines are found along the edge ridges, and divide the image into closed and connected regions/cells (basins). Thus there is a high degree of overlap between the watershed lines of a gradient (magnitude) image and the result after non-maximum suppression. Based on this relationship, the WT of the combined input gradient map (from Section 4.2) is used to provide a meaningful completion guide to connect any broken contours in the $\mathrm{tCSM_b}$.

In Fig. 10(a), we show the combined input gradient map corresponding to the ROIs in the second row of Fig. 9. Fig. 10(b) and (c) show the WT of this image overlaid on the thermal and visible ROIs, respectively. In Fig. 10(d), we show the corresponding $\mathrm{tCSM_b}$ overlaid on the WT lines. This image shows the overlap between the $\mathrm{tCSM_b}$

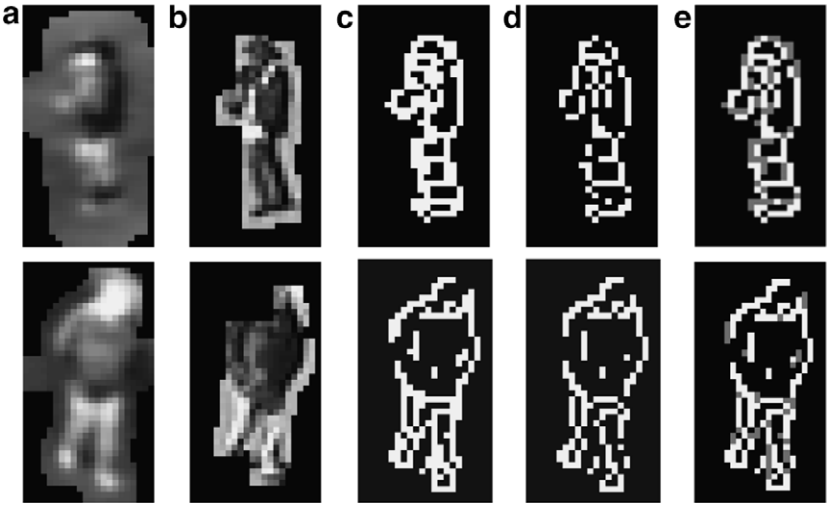


Fig. 9. Fused binary contours. (a) Thermal ROI. (b) Visible/intensity ROI. (c) Fused binary contours before alignment. (d) Fused binary contours after alignment. (e) Contours from (d) (white) overlaid over contours from (c) (gray).

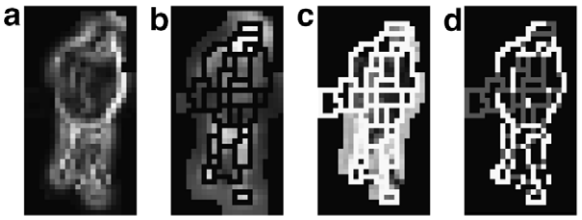


Fig. 10. Watershed analysis. (a) Combined input gradient map. (b) Overlay of watershed lines on thermal ROI and (c) visible ROI. (d) Superposition of $tCSM_b$ on watershed lines.

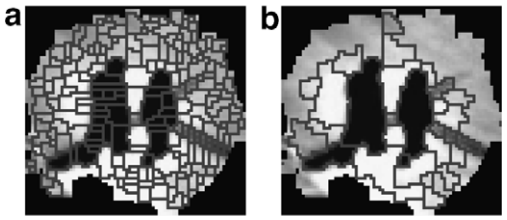


Fig. 11. Basin merging. (a) Original WT overlaid on ROI. (b) Merged WT overlaid on ROI.

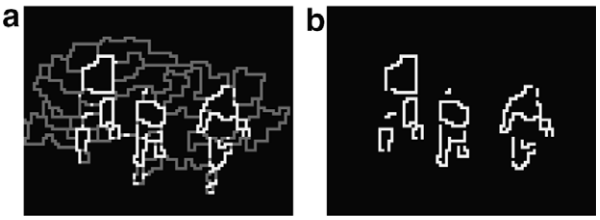


Fig. 12. Contour validation. (a) $tCSM_b$ overlaid on merged watershed lines. (b) $tCSM_b$ after contour validation.

and the WT lines, and also the suitability of the watershed lines as appropriate completion guides for the broken contours of the $tCSM_b$.

4.3.1. Valid contour selection

To eliminate small stray contour fragments present in the $tCSM_b$ that may harm the completion/closing process, a coarser segmentation of the ROI is obtained using a basin-merging algorithm on the watershed partition. The basin merging algorithm uses the Student's *t*-test with a confidence threshold of 99% to determine whether the pixels for two adjacent basins in the ROI are similar (merge) or significantly different (do not merge). Starting from the two most similar basins, pairs of basins are merged until no two neighboring basins pass the similarity test. The merged version of the WT generates a lower resolution segmentation of the ROI. We shown an example before and after basin merging in Fig. 11. Other merge algorithms could also be applied [33,26].

This coarser resolution WT is used to validate the contour segments of the thresholded $tCSM_b$ by eliminating any small noisy fragments that might exist. Based on the merged WT, the binary $tCSM_b$ is partitioned into distinct segments that divide pairs of adjacent basins. A $tCSM_b$ segment is considered valid only if its length is at least 50% of the length of the WT border separating the two basins. If a segment is deemed invalid, its pixels are removed from the thresholded $tCSM_b$. The intuition behind the process is that, at least half of the boundary between two neighboring regions must be reinforced, otherwise the $tCSM_b$ pixels on the boundary are likely to be noise. We show a threhsolded $tCSM_b$ overlaid on the merged WT in Fig. 12(a) and show the result after the validation process in Fig. 12(b). Notice that several small fragments are removed after the validation process. The merged WT is used only for this step, to validate the contours in the $tCSM_b$, and the remaining completion/closing processes employ the original WT.

4.3.2. Contour completion

The first part of the method attempts to complete any contour gaps using the watershed lines as plausible connection pathways for the $tCSM_b$. Each "loose" endpoint of the contour segments (found using 3×3 neighborhood analysis) is forced to grow outward along the watershed lines until another contour point is reached. To find the optimal path, the A* search algorithm [42] is employed such that the expected cost *through* the current pixel location to reach another contour point is minimized. The Euclidean distance from the current pixel location to the location of remaining thresholded tCSM contour pixels is used as the heuristic cost function. Each gap completion search uses only the original contour pixels (not including any new path points) so that the order of the gap completion does not influence the final result. Again, the valid search paths are restricted to only the watershed lines. Further details of the approach are provided in [13]. We show a $tCSM_b$ in Fig. 13(a) and the completion result in Fig. 13(b). The un-completed contour in the foot region of the middle person and the other un-closed contour segments (shown in white) will be addressed in the following closing operation.

4.3.3. Contour closing

Next all those contours not part of a closed loop are identified (by region growing along the contours). Such contours either have a remaining external "loose" end (not completed in the previous step) or connect other closed loops. Given an un-closed contour, the nearest pair of points that lie on closed loops connected by this contour are chosen. Any other path connecting these points will close the region. As in the previous step, the A* algorithm limited to the watershed lines is used to find such a path. In order to ensure that only a minimum number of new contour pixels are added to the figure by the new path, no penalty (step cost) is assigned for moving along existing contour pixels during the A* search. The process of contour closing is repeated until no new pixels are added to the image between successive iterations. We show example closing results in Fig. 13(c).

After the completion and closing procedures, a simple flood-fill operation is employed to create the silhouettes.

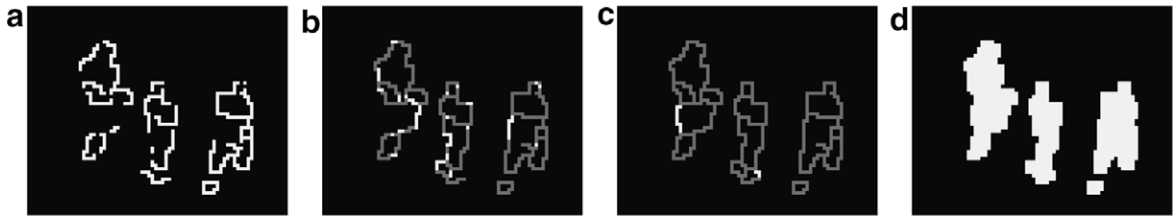


Fig. 13. Contour completion, closing, and flood-filling. (a) Original $tCSM_b$. (b) Completed contour result (white lines are new paths). (c) Closed result of (b) (white lines are new paths). (d) Flood-filled silhouettes.

We present the final flood-filled silhouettes for the closing result of Fig. 13(c) in Fig. 13(d).

5. Stage III: high-level processing

In this stage, we weight each resulting silhouette with a contrast value, $\mathscr{C}$, representative of how distinct that region is from the background scene in either domain. For each silhouette, we first compute the ratio of the maximum input-background intensity difference within the silhouette region to the full intensity range of the background image BG, in both the thermal and visible domains. We then assign the higher of these two values as the contrast value of the silhouette.

$$\mathscr{C}(\mathrm{sil}) = \max_{i\in\{T,V\}} \left(\frac{\| \max(I^i(\mathrm{sil})) - \max(\mathrm{BG}^i(\mathrm{sil})) \|}{\max(\mathrm{BG}^i) - \min(\mathrm{BG}^i)} \right) \qquad (10)$$

where sil represents a particular silhouette region detected in the input image I. The contrast (or confidence) level ($\mathscr{C}$) for each silhouette could also be determined by more robust methods, such as estimating the divergence of each silhouette distribution from the background distribution. However, due to the large dynamic range of our images, this simple method was effective. The contrast value provides the flexibility to a human operator (e.g., for human-in-the-loop surveillance monitoring) to select/show only the most confident detections. A user-selected threshold on $\mathscr{C}$ could easily be used to remove any minimal-contrast (noise) regions.

To further improve detection results, we also make use of limited temporal information. A 3-tap temporal median filter is applied to blobs across frames $\{I_{t-1}, I_t, I_{t+1}\}$ to ensure that sporadic detections are eliminated. The filter preserves only those silhouette regions in the current frame that were also detected at least once in a local temporal window. Though our approach is simple, it is effective and does not burden the pipeline with excessive memory/computational requirements. A deeper median filter (say 5- or 7-tap), or more complex temporal consistency algorithms, along with grouping/clustering of small silhouette blobs, could also be employed. We also note that similar temporal information could be used to better facilitate the identification of ROIs in Stage I of the algorithm.

In Fig. 14, we show the results of this processing stage. Fig. 14(c) shows the initial silhouette regions detected for the pair of images shown in Fig. 14(a) and (b). In Fig. 14(d), we show the result of removing sporadic inconsistent detections using the median filter and assigning contrast values (darker silhouettes denote detections with lower confidence).

6. Experiments

To examine our contour-based fusion approach, we tested our method with six challenging thermal/color video sequence pairs recorded from two different locations at different times-of-day, with different camera gain and level settings. We also analyzed our method quantitatively with a comparison to alternate approaches. The thermal sequences were captured using a Raytheon 300D ferroelectric BST thermal sensor core, and a Sony TRV87 Handycam was used to capture the color sequences. The image sizes were half-resolution at 320×240. The number of frames in each sequence is Sequence-1:2107, Sequence-2:1201, Sequence-3:3399, Sequence-4:3011, Sequence-5:4061, and Sequence-6:3303. Example images from this dataset, one from each sequence, are shown in the top two rows of Fig. 15(a)–(f). The sequences were recorded on the Ohio State University campus during the months of February and March 2005, and show several people, some in groups, moving through the scene. Sequences 1, 2, and 3 contain regions of dark shadows cast by the buildings in the background. There are also frequent (and drastic) illumination changes across the scene. The images of Sequences 4, 5, and 6 were captured on a cloudy day, with fairly constant illumination and soft/diffuse shadows. To incorporate variations in the thermal domain, the gain/level settings on the thermal camera were varied across the sequences.

6.1. Qualitative analysis

Examples of silhouettes extracted using the proposed fusion-based background-subtraction method are shown in the bottom row of Fig. 15(a)–(f). To demonstrate the generality and applicability of our approach, the silhouettes were obtained using the same parameter/threshold settings for all sequences (squared Mahalanobis distance thresholds of 10, 15, and 4 for the thermal, luminosity, and color channels, respectively). The results demonstrate that the algorithm is able to generate reasonable silhouette shapes even when objects are difficult to discern individually

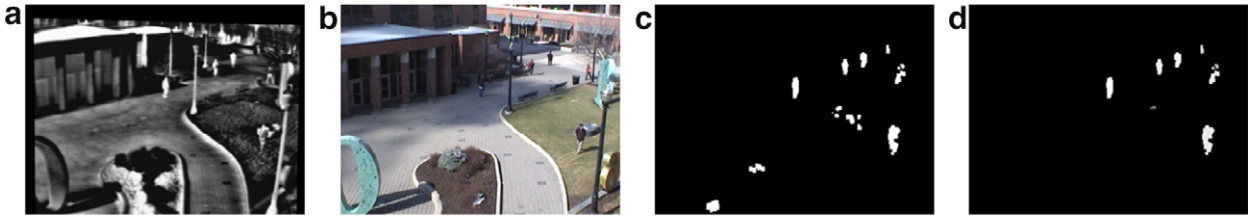


Fig. 14. High-level processing. (a) Input thermal image. (b) Input visible image. (c) Detected silhouette regions. (d) Silhouette regions after high-level (post-) processing.

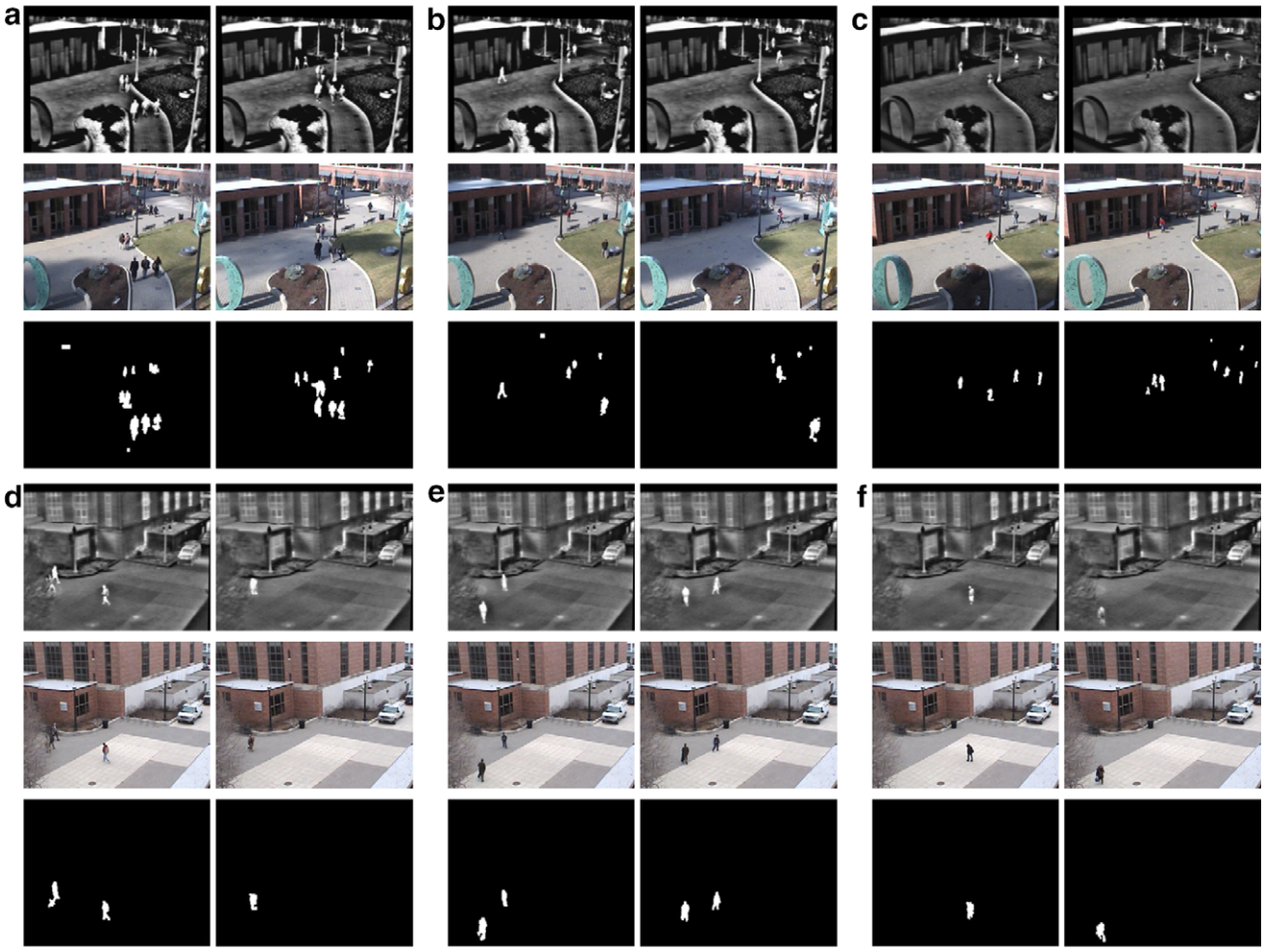


Fig. 15. Visual inspection of detection results of the proposed approach across different images and scenarios. (a) Sequence-1. (b) Sequence-2. (c) Sequence-3. (d) Sequence-4. (e) Sequence-5. (f) Sequence-6.

in the two input channels. In Sequences 1, 2, and 3 the algorithm is unaffected by the sudden changes in illumination in the visible domain. It is also able to ignore shadows of people in the scene. The variation in thermal gain settings across these sequences did not pose any problems. Sequences 4, 5, and 6 contained more stable illumination and diffused shadows. The thermal domain contained images that showed a slight halo artifact. The corresponding silhouettes extracted show that our algorithm was effective in handling these conditions. Examining the resulting silhouette images across all 6 sequences, we also see that the contour completion and closing method is able to effectively separate multiple people within a single ROI.

Some visible and thermal regions containing people are shown in the three rows of Fig. 16(a) and (b), respectively. In the visible (Fig. 16(a)) and thermal (Fig. 16(b)) image pair shown in the top row, we see an example of the contrast between the two domains. The visible image shows a challenging case where the person appears dark and is in shadow. In the corresponding thermal image, the person regions appear white hot, possibly due to the lower temperatures in the shade. In the middle row, different person

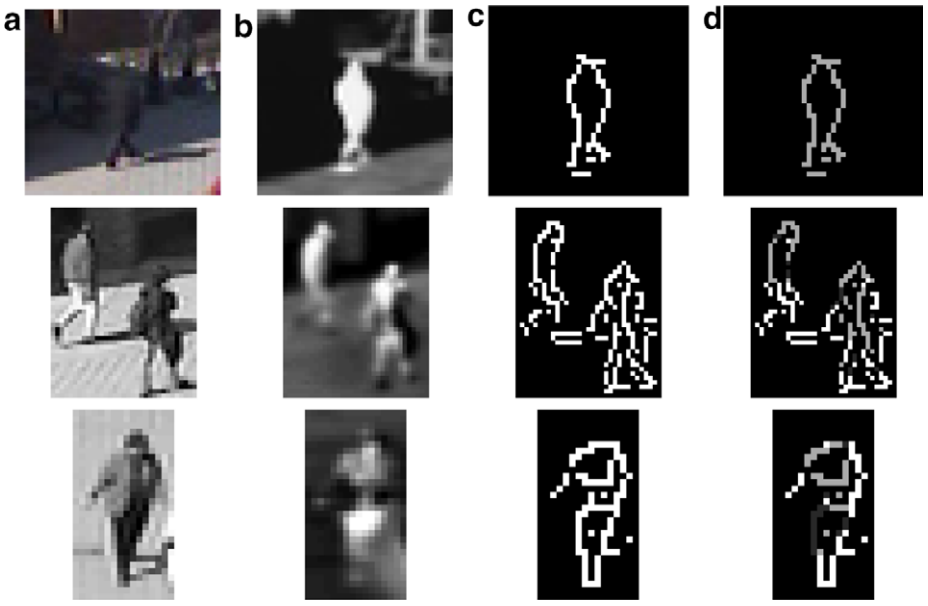


Fig. 16. Three example person regions. (a) Visible domain. (b) Thermal domain. (c) Final fused contours. (d) Contours present in both domains (dark gray), only in thermal domain (gray), and only in visible domain (white).

regions appear at varying contrasts from the background in both the visible and thermal domains. Observe, however, that regions of high or low contrasts in one domain do not correspond to similar contrasts in the other domain. In the third row, we see that the person regions are more clearly delineated in the visible domain than in the thermal domain.

In Fig. 16(c), we show the results of the proposed approach, before contour completion, for the three examples. The extracted contours provide a reasonable, albeit incomplete, trace of the boundary of the person regions shown in Fig. 16(a) and (b). In Fig. 16(d), the contours have been color-coded to show the contribution of either domain to the final fused contour image. Contours in white are those obtained from the visible domain, those in gray are obtained from the thermal domain, and those in dark gray are common to both domains. The contour image in the top row of Fig. 16(d) is composed of information from the thermal domain alone. This is because the visible ROI, extracted from the corresponding image in Fig. 16(a), contained no usable contour information. The contours in the second row of Fig. 16(d) are obtained almost equally from the visible and thermal domains. Portions of the torso, that are not clearly distinguishable in the visible image, are outlined by contours from the thermal domain, while the leg regions are mostly outlined by contours from the visible domain. In the third row, we see an example where the contours from the visible domain provide most of the shape information of the person region. Here, the thermal domain contributes contours that are mainly internal to the person regions.

In order to determine the usefulness of our fusion approach, and to enable comparison with other algorithms, we use object-detection performance as a common yardstick. To quantitatively measure object-detection performance, we obtained a manual segmentation of the person regions in 60 image-pairs from our dataset (~10 image-pairs spanning each of the 6 sequences). For each of the 60 image-pairs, three people hand-segmented the person regions (silhouettes), in both the thermal and visible domains. Results of the hand-segmentation of each pair of images by each person were combined using an element-wise logical OR operation to obtain the final manual silhouette image. The median silhouette images across the 3 participants were used in the algorithm evaluation. Examples, one from each sequence, are shown in the top row of Fig. 17.

Using the manually segmented images, we performed two sets of experiments. The first experiment was used to ascertain if object detection performance is improved using our fusion approach, as opposed to using either of the two sensors independently. The second set of experiments compared our approach with other potential methods for fusion-based background-subtraction. These methods include both low-level (pixel-based) and high-level (decision-based) schemes, which together with our mid-level (gradient-based) approach, make up a rich testbed of fusion techniques.

6.2. Experiment 1: fusion vs. independent sensors

The aim of this experiment is to demonstrate the efficacy of our fusion algorithm for object detection. As a measure of efficacy, we computed the improvement in object-detection that our algorithm achieved by using thermal and visible imagery together, over using either of the sensors independently. The performance of our algorithm was thus examined under three different input scenarios, thermal

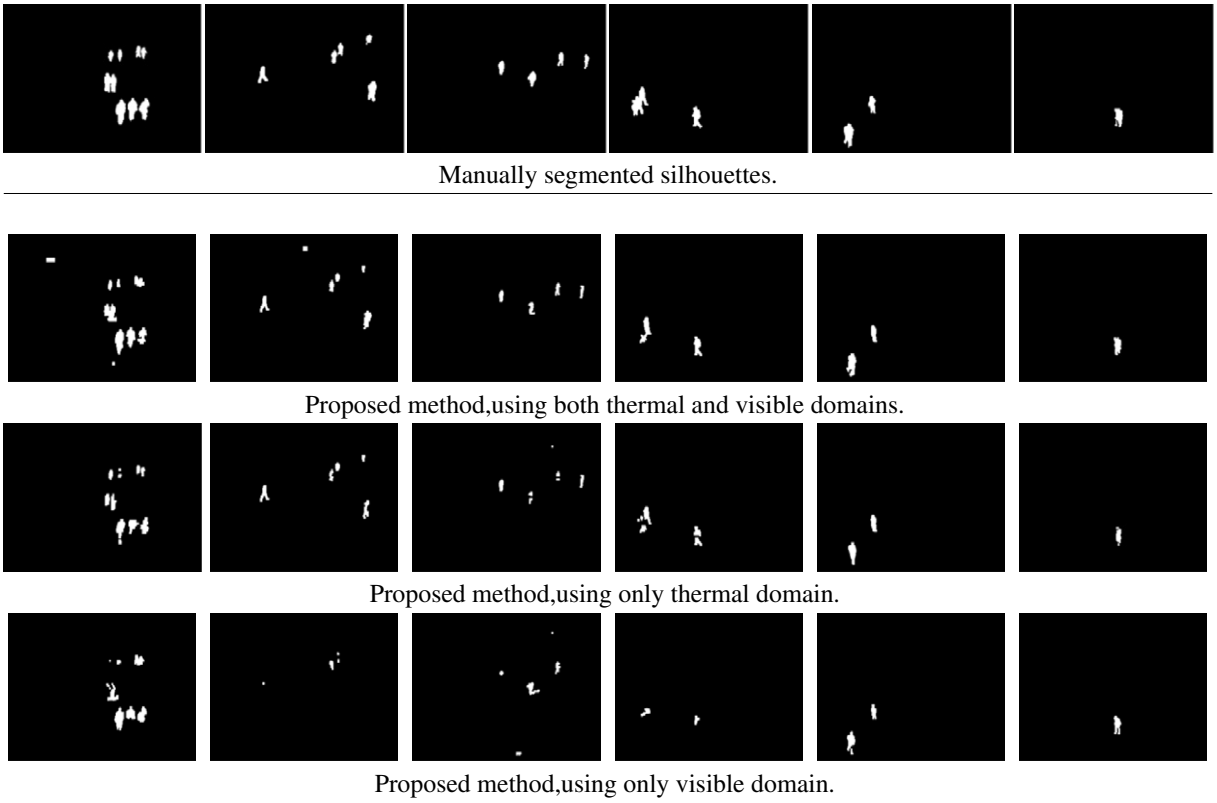


Fig. 17. Visual comparison of detection results of the proposed approach for across different images and input scenarios.

and visible imagery, only thermal imagery, and only visible imagery. We measured *Recall* and *Precision* for each scenario using the set of 60 manually labeled images as ground truth. Recall (or Sensitivity) refers to the fraction of object pixels that are correctly detected by the algorithm, while Precision (or Positive Predictive Value) represents the fraction of detections that are in fact object pixels. In all our experiments we use the *F*-measure [41], which is the harmonic mean of Precision and Recall, as an indicator of the quality of performance

$$F = \frac{2PR}{P+R} \tag{11}$$

In a Precision–Recall graph, higher *F*-measures correspond to points closer to $(1,1)$, representing maximum Recall and Precision.

In this experiment, we compared the three input scenarios over a large number of parameter settings. The parameters we varied were the 'internal' background-subtraction thresholds (corresponding to the squared Mahalanobis distance thresholds for the thermal, luminosity, and color channels), and the 'external' threshold (on the contrast value $\mathscr{C}$). The internal thresholds (e.g., as used in Eq. 5) were varied in the following ranges: $[5,6,\ldots,12]$ for the thermal, $[10,11,\ldots,16]$ for the luminosity, and $[2,4,\ldots,8]$ for the color components. These values resulted in 245 possible parameter settings. Other statistics-based methods could also be used to determine the optimal thresholds. For each internal parameter setting, the external parameter was varied between 0 and 1, and assigned the value giving the highest *F*-measure of Precision and Recall. Examples of detection results for the three input scenarios (thermal and visible, thermal alone, and visible alone) are shown in rows 2, 3, and 4 of Fig. 17. The corresponding ground truth silhouettes for these examples are shown in the top row of Fig. 17. In Tables 1–3, we show comparisons between Recall, Precision, and the *F*-measure for the three scenarios for a particular set of internal parameters (thresholds for thermal, luminosity and color set at 10, 15, and 9, respectively). The external threshold was set independently for each input scenario, such that the *F*-measure was maximized.

From column 2 of Table 1 we see that, using only the thermal domain, the Precision values for all sequences are close to 1. This shows that the proposed algorithm is effective in extracting object silhouettes despite issues

Table 1
Comparison of *Precision* values

Sequence	T and V	T	V	% over T	% over V
Seq-1	0.914	0.938	0.459	−2.56	99.13
Seq-2	0.874	0.926	0.215	−5.62	306.52
Seq-3	0.905	0.922	0.773	−1.84	17.07
Seq-4	0.955	0.958	0.983	−0.31	−2.85
Seq-5	0.957	0.966	0.954	−0.93	0.31
Seq-6	0.937	0.941	0.952	−0.43	−1.57
Overall	0.916	0.939	0.498	−2.45	83.94

Table 2
Comparison of *Recall* values

Sequence	T and V	T	V	% over T	% over V
Seq-1	0.714	0.607	0.313	17.63	128.12
Seq-2	0.719	0.6668	0.130	7.64	453.07
Seq-3	0.655	0.569	0.218	15.11	200.46
Seq-4	0.734	0.718	0.122	2.23	501.63
Seq-5	0.809	0.777	0.148	4.12	446.62
Seq-6	0.78	0.663	0.439	17.65	77.67
Overall	0.722	0.645	0.233	11.94	209.87

Table 3
Comparison of *F*-measure

Sequence	T and V	T	V	% over T	% over V
Seq-1	0.802	0.737	0.372	8.774	115.402
Seq-2	0.789	0.776	0.162	1.654	386.925
Seq-3	0.760	0.704	0.340	7.994	123.462
Seq-4	0.830	0.821	0.217	1.124	282.401
Seq-5	0.877	0.861	0.256	1.805	242.169
Seq-6	0.851	0.778	0.601	9.437	41.674
Overall	0.808	0.765	0.317	5.596	154.361

including the halo artifact and the inversion of polarity. This also demonstrates that the algorithm is capable of generating good object detection performance when the visible domain provides little or no information, such as in nighttime environments. In fact, as we show in [13], even better performance could be achieved using cooler nighttime sequences (or hotter daytime sequences), since the thermal properties of person regions would be significantly different from that of the environment. The addition of information from the visible domain does not improve the Precision further, and in fact results in a marginal percentage degradation of Precision, as shown in column 4 of the table. The relatively poor performance obtained from using only the visible sensor is to be expected given our simple background model and the presence of shadows and drastic illumination changes occurring throughout the sequences. Without the use of a thermal mask (Section 3.2), extracting ROIs in the visible domain using the same set of thresholds for all 6 sequences results in a large number of spurious detections (see Fig. 4). In the visible domain (column 3), Sequences 1–3 have low Precision due to presence of sharp shadows and drastic illumination changes, while Sequences 4–6 have high Precision due to more benign conditions. Incorporating information from the thermal domain results in a large percentage improvement for Sequences 1–3, while the Precision drops slightly for sequences 4 and 6, as shown in column 5. Overall, using our approach, fusion of thermal and visible imagery results in a very small (~2%) percentage degradation of Precision compared to using only the thermal domain, and a large (~84%) percentage improvement when compared to using only the visible domain.

The benefits of fusing information becomes apparent while comparing Recall values, shown in Table 2. Examining columns 4 and 5, we see that, for all six sequences, fusion of thermal and visible imagery results in large percentage improvements over using either domain independently. Overall, there is almost a 12% gain in Recall over the thermal domain, and the Recall of the visible domain is more than doubled.

Table 3 shows the maximum *F*-measure of the Precision and Recall for each input scenario. Column 4 shows that our fusion-approach generates measurable performance gains over using only thermal imagery for all sequences. The positive change in overall performance shows that the improvement in Recall brought about by fusion with the visible domain outweighs the marginal loss incurred in Precision. Further, we see from column 5 that our approach improves results dramatically by using both the thermal and visible domains over using only the visible domain.

The performance gains shown in Tables 1 and 2 are not peculiar to the aforementioned threshold values. In Fig. 18(a)–(c), we show Precision–Recall plots for three different internal parameter settings. The thermal, luminosity, and color squared Mahalanobis distance thresholds (e.g., as used in Eq. 5) are [8, 13, 6], [10, 15, 10], and [9, 14, 8] for Fig. 18(a)–(c), respectively. Each plot shows the inter-dependence of Precision and Recall as the external threshold (on contrast $\mathscr{C}$) is varied. The highlighted points correspond to the thresholds that yield the highest *F*-measure.

Examining the three curves plotted in each of the figures, we see that for all the parameter settings shown, the detection results obtained by fusing thermal and visible imagery using our method are superior to using either kind of imagery individually. In Table 4, we show the highest *F*-measure obtained for seven of the many internal parameter settings we tested. As shown by the results in the table, our algorithm generated better detection performances when using both thermal and visible imagery for the majority of parameter settings. The last column in the table shows the standard-deviation of the performance of each input scenario across the different parameter settings. A higher standard deviation signifies a higher sensitivity to the threshold values, and hence lower consistency or robustness. As expected, the visible domain is the least consistent of the three cases. The thermal domain, on the other hand, shows the most consistent results across the various settings. In comparison, the fusion of the two domains by the proposed algorithm produces fairly consistent results. In spite of the highly sensitive visible domain, our method is able to exploit the robustness of thermal imagery to produce results that are closer in consistency to the thermal, than the visible, domain.

6.3. Experiment 2: comparison with other methods

In this experiment we compare the proposed algorithm with other potential fusion-based background-subtraction methods. Our mid-level algorithm combines information from the two domains at the gradient (or contour) level.

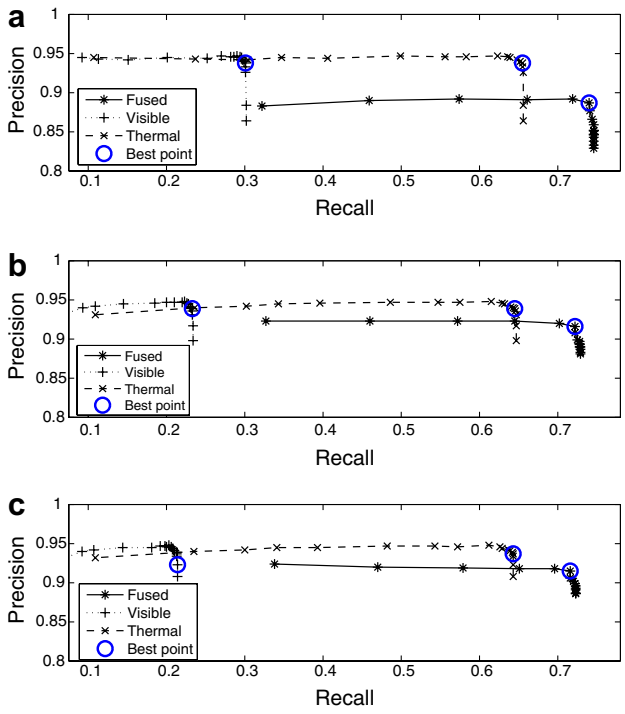


Fig. 18. Precision–Recall graphs for three different threshold settings for thermal, luminosity, and color channels. (a) [8,13,6]. (b) [10,15,10]. (c) [9,14,8].

In order to encompass the broadest range of techniques, we compare our method with lower- and higher-level fusion algorithms. The two low-level (pixel-level) fusion techniques we use are simple averaging and the PCA-based technique proposed in [8]. These techniques generate fused imagery which can be then be used as the input stream for background-subtraction algorithms. The high-level (decision-level) technique we employ is the logical OR-ing of binary silhouettes, obtained *after* background-subtraction is performed independently in both domains. In order to determine which background-subtraction techniques to use for comparison, we note that these can be roughly categorized as either parametric or non-parametric approaches. From among the former, we consider Gaussian-based modeling methods and choose the single Gaussian model [50] for comparison, since it is a simple technique that is still widely used, and because we employ this technique to generate the initial ROIs in our method. The more complex Mixture of Gaussians [45] model could also be used, but would be equally ineffective against halos and object shadows found in thermal and visible imagery (halos and shadows co-occur temporally and spatially with foreground objects). From the non-parametric class of background-subtraction algorithms, we use the code-book based approach proposed in [24]. This recent technique has been shown to generate comparable or better results than the popular Mixture of Gaussians approach. The full set of competing methods can be enumerated as follows:

1. *PCA-based fusion + Gaussian back-sub:* Fusion is performed by computing a region-wise weighted average of the input sensors. The weights for each circular region are determined using PCA of the pixel-intensities of the input images (for details see [8]). We use the method to fuse each color component of the visible domain with the thermal channel resulting in a fused image stream with three components. Background-subtraction is performed using a mean-covariance (3-dimensional) Gaussian model. Based on visual inspection of the fusion results, we set the standard deviation of the low-pass filter used in the technique to 2, and the radius of the circular regions to 5.
2. *Average + Gaussian back-sub:* Pixel-wise averages are computed for each color channel with the corresponding thermal channel. As in Method 1, a 3-dimensional mean-covariance Gaussian model is used for background-subtraction.
3. *Proposed algorithm:* As described in Sections 3–5.
4. *Gaussian back-sub in each domain + OR-ing of silhouettes:* 1-d and 3-d Gaussian background models are built for the thermal and visible domains, respectively. The background subtraction results are combined using a pixel-wise logical OR operation.
5. *Code-book back-sub in each domain + OR-ing of silhouettes:* The code-book technique [24] is used to build the background models for the thermal and visible domains. As in Method 4, the background-subtraction results are combined using pixel-wise logical OR.

For Methods 1, 2, and 4 we also employ the same shadow-removal approach described in Section 3.2, where pixels darker in luminosity and similar in chrominance with the background are discarded. To ensure a fair comparison, contrast values ($\mathscr{C}$) were computed for silhouettes generated by each method. Further, the internal parameters for all the methods were set to guarantee the best performance on the testset. In Methods 1–4, each

Table 4
Overall *F*-measure values for different threshold settings for thermal, luminosity, and color

Method	Setting-1 [5,10,4]	Setting-2 [6,11,5]	Setting-3 [7,12,6]	Setting-4 [8,13,7]	Setting-5 [9,14,8]	Setting-6 [10,15,9]	Setting-7 [11,16,10]	σ
Thermal and visible	0.781	0.792	0.798	0.807	0.804	0.808	0.803	0.0096
Thermal only	0.772	0.773	0.771	0.771	0.769	0.765	0.763	0.0040
Visible only	0.659	0.575	0.513	0.456	0.405	0.373	0.374	0.1087

background-subtraction threshold was varied over a wide range. For each of the large number of resulting parameter sets, we plotted the Precision–Recall curve by varying the threshold on $\mathscr{C}$. The best F-measure of the curve was used as the summary statistic for the particular parameter set. For each method, the parameter settings with the highest summary statistic were then chosen as the internal parameters. It should be noted that for these methods, the exact same parameters were used for all the 6 sequences and for both the thermal and visible input channels. For Method 5, the several parameters pertaining to the code-book technique [24] ($\epsilon 1$, $\epsilon 2$, α, β, K, Period, parameters used for post-processing and extreme point detection, and parameters used for the layered background model updating) were handpicked by the authors of [24] for each of the 6 different sequences and fine-tuned individually for the thermal and visible domains. Thus, for Method 5, only the threshold on $\mathscr{C}$ was chosen based on the highest F-measure.

In Fig. 19, we show the best Precision–Recall curves for the 5 methods. The curves are generated by varying the threshold on $\mathscr{C}$, and the point with the highest F-measure is highlighted. For each method, the Precision and Recall values corresponding to this point are shown in Table 5. The last column of the table shows the Precision, Recall and F-measure computed over the entire testset of 60 images. Examples of detection results of the five methods are shown in rows 2–6 of Fig. 20. To enable a visual comparison, the corresponding ground-truth data is shown in row 1 of the figure. From Table 5, we see that among the low-level techniques, Method 1 has a slightly better F-measure than Method 2. Both methods have low detection rates, and less than 65% of the pixels detected lie on target. This shows that while these methods might be useful as visualization aids, they are not very effective at facilitating automatic object detection via background-subtraction. We see that the highest Recall value is obtained by Method 4. This is because background-subtraction in the thermal domain is very successful at accounting for almost all of the person pixels present in the testset. However, this technique is unable to effectively handle the halo artifact common in thermal imagery. As a result, a large number of pixels *around* the objects are incorrectly detected. In the visible domain, the presence of shadows and illumination changes causes further problems for the background-subtraction technique (as such statistical techniques do not directly model shadow/illumination properties as in the code-book method), resulting in a number of false alarms. Since silhouettes from both domains are combined using a simple logical OR, the false detections from both domains contribute to the low Precision values for Method 4. The code-book technique for background-subtraction used in Method 5 generates more balanced results. The higher Precision indicates that the technique is able to deal with issues such as halos and shadows more effectively than the statistical background-subtraction approach. However, this comes at the cost of Recall, which is lower than that of Method 4. The better overall performance can also be attributed to the fact that the parameters used in the code-book technique were fine-tuned for each sequence for the thermal and visible domains.

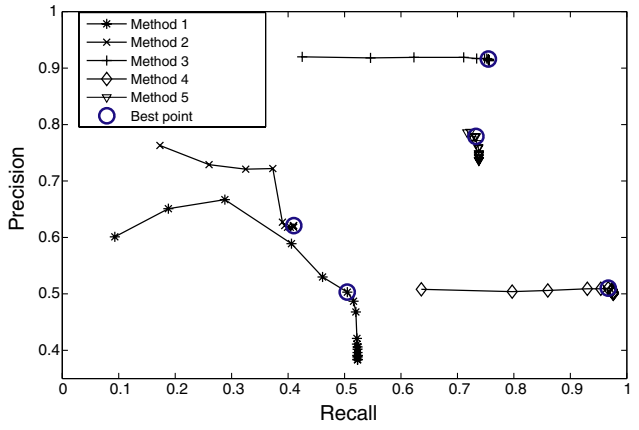


Fig. 19. Best Precision–Recall plots for five different methods.

Among all the methods, the highest F-measure is achieved by the proposed algorithm (Method 3). It has a considerably higher Precision than the competing methods, showing that it is best able to deal with the challenges of both domains, namely halos, shadows, and illumination changes. At the same time, it also has a Recall rate of more than 75%, second only to Method 4, which, as discussed earlier, has a large false positive rate.

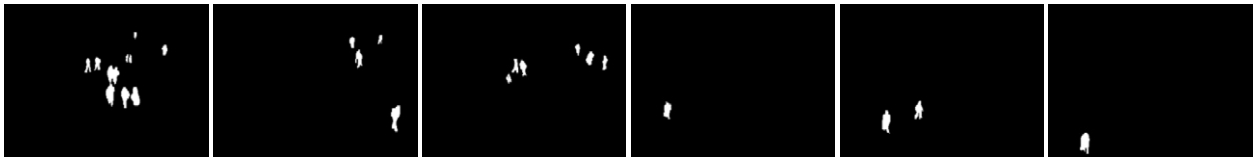
6.4. Discussion of experiments

A number of salient properties of our algorithm are brought forth in these experiments. In Experiment 1, we start by showing that our algorithm is able to exploit the presence of two complementary input modalities to improve object detection performance. The overall detection results (rated using the F-measure) when both thermal and visible imagery are used are always better than when only one domain is present (see Table 3). We regard this as a fundamental requirement of a fusion technique, whereby fusion of a number of sensors does not degrade performance over using any one of the input modalities alone. Another useful quality of our algorithm demonstrated by this experiment is the graceful degradation of performance as any one of the input sensors becomes unavailable. Next, we show that the improvement brought about by the proposed mid-level fusion technique is robust to large variations in parameter settings. In spite of the rapidly changing nature of visible imagery, our algorithm is able to exploit the stability of thermal imagery to generate results that are reasonably consistent over a wide parameter space (see Table 4).

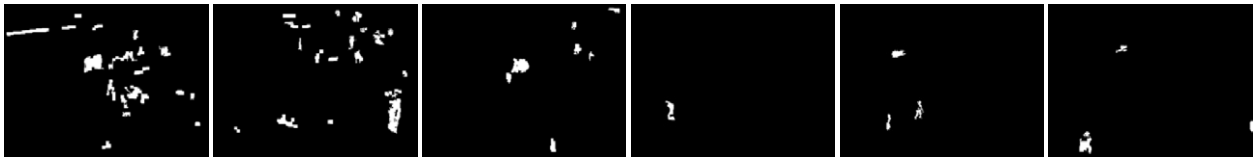
As discussed earlier, fusion can be performed at several stages in an object-detection pipeline. Experiment 2 compares our mid-level, contour-based technique against a wide variety of methods, including low- and high-level

Table 5
Comparison of Precision (P) and Recall (R) values of different fusion methods across different sequences

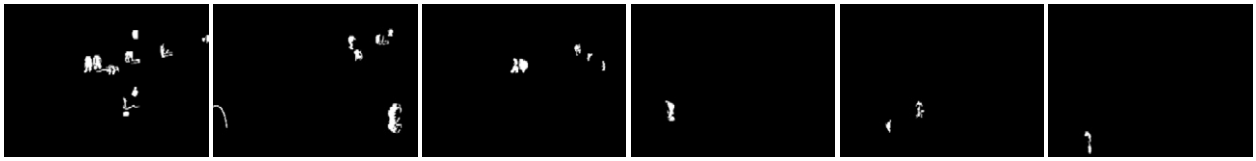
	Method		Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	Overall	
Low-level	Method 1	R	0.454	0.649	0.489	0.518	0.440	0.518	0.505	$F = 0.504$
		P	0.519	0.402	0.498	0.694	0.601	0.613	0.503	
	Method 2	R	0.336	0.586	0.422	0.547	0.279	0.344	0.410	$F = 0.491$
		P	0.747	0.442	0.562	0.895	0.836	0.771	0.613	
Mid-level	**Method 3**	R	0.756	0.754	0.683	0.759	0.823	0.814	0.755	$\boldsymbol{F = 0.828}$
		P	0.908	0.890	0.900	0.958	0.965	0.931	0.916	
High-level	Method 4	R	0.883	0.928	0.908	0.882	0.957	0.965	0.910	$F = 0.748$
		P	0.688	0.518	0.590	0.712	0.748	0.668	0.635	
	Method 5	R	0.772	0.561	0.543	0.925	0.932	0.914	0.733	$F = 0.755$
		P	0.747	0.568	0.915	0.910	0.909	0.915	0.779	



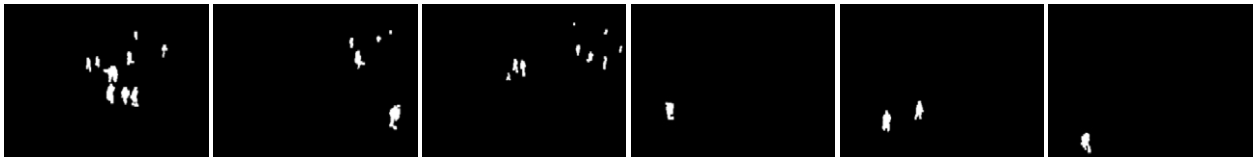
Manually segmented silhouettes.



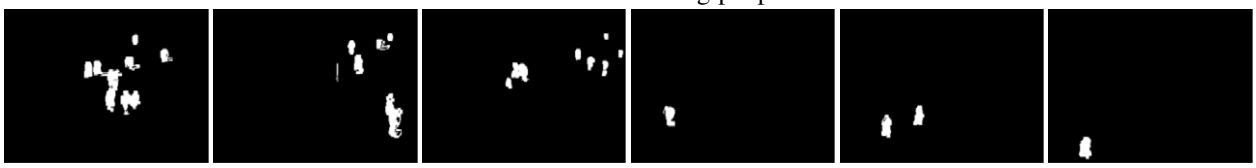
Method 1: Low-level fusion using PCA-based method.



Method 2: Low-level fusion using pixel averages.



Method 3: Mid-level fusion using proposed method.



Method 4: Decision level fusion using Gaussian back-sub.

Method 5: Decision level fusion using code-book back-sub.

Fig. 20. Visual comparison of detection results of five different methods across different images and scenarios.

schemes. This empirical evaluation shows that, for object detection using thermal and visible imagery, our contour-based approach avoids many of the pit-falls of low- and high-level techniques (see Table 5). A mid-level fusion strategy is best suited to exploit the favorable characteristics of both input modalities and protect against the undesirable artifacts. The salient contour features we extract enables us to focus on high contrasts in either domain while

avoiding difficulties arising due to halos and diffused shadows. Further, the lower-level interaction (see Fig. 2) between the background-subtraction results of the two domains enables us to overcome issues in the visible domain pertaining to sudden illumination changes (as discussed in Section 3.2).

6.5. Computational cost

Since our algorithm processes each ROI identified within an image individually, the frame-rate is heavily dependent on the number of ROIs found in each image. The burden of finding the ROIs in either domain is dependent on the background subtraction technique used. As discussed (Section 3.2), satisfactory background-subtraction results can be obtained from the relatively stable, single channel, thermal domain, using only simple techniques (such as the mean-variance model). Background-subtraction in the visible domain is inherently more expensive due to the presence of three color channels (RGB). However, in our method, once the thermal ROIs are obtained, background-subtraction in the visible domain is performed only within the thermal ROIs, and not over the entire image, thereby providing a reduction in computation cost.

Once the ROIs have been identified, the most expensive components of the algorithm are the search and validation methods used in contour completion and closing. The expensive Watershed Transform [49], used in contour completion and closing, however does not contribute significantly to the overall processing time since it is only applied in small ROIs (rather than the entire image). Using un-optimized Matlab code on a 2.8 GHz Pentium 4 computer, we experienced typical processing times in the range of 0.8–5.5 s per ROI to generate the final fused silhouettes depending on the complexity of the ROI (number of people, modality).

7. Summary

We presented a new contour-based method for combining information from visible and thermal sensors to enable persistent background-subtraction in urban scenarios. Our approach handles the problems typically associated with thermal imagery produced by common ferroelectric BST sensors such as halo artifacts and uncalibrated polarity switches, using the method initially proposed in [11]. The problems associated with color imagery, namely shadows and illumination changes, are handled using standard techniques that rely on the intensity and chromaticity content.

Our approach first uses statistical background-subtraction in the thermal domain to identify local regions-of-interest containing the foreground object and the surrounding halo. Color and intensity information is then used within these regions to extract the corresponding regions-of-interest (without shadows) in the visible domain. The input and background gradients within each region are then combined into a Contour Saliency Map (CSM). The CSM is thinned using a non-maximum suppression mask of the individual input gradients. The most salient contours are then selected using a thresholding strategy based on competitive clustering. The binary contours from corresponding regions of the thermal and visible image are then combined and thinned using the input gradient information from both sensors. Any broken contour fragments are completed and closed using a watershed-constrained A^* search strategy and flood-filled to produce silhouettes. Each silhouette region is then assigned a confidence value based on its contrast with the background. Lastly, noisy silhouette regions are eliminated in a post-processing stage.

Experiments were conducted with six challenging thermal/color video sequences recorded from different locations and at different times-of-day. Our algorithm generated promising results using a single set of parameters/thresholds across all 6 sequences. We used a set of 60 manually segmented images to perform a thorough quantitative analysis based on the Precision and Recall values of the detected silhouette regions. We first demonstrated that our algorithm is able to effectively exploit information from the thermal and visible domains to improve object detection performance. We showed that our algorithm generates improved object detection performance when using thermal and visible imagery together, over using either domain independently. The fact that these performance gains were obtained over a wide range of parameter settings also demonstrated the robustness (consistency) of our approach. In another experiment we compared our mid-level fusion algorithm with other potential low- and high-level fusion methods for background-subtraction. Our algorithm consistently generated the best detection performance (in terms of the F-measure), providing empirical evidence for the superiority of our algorithm.

To further improve our results, we plan to include motion information into the saliency map and employ shaped-based models for better figure completion. We also recognize the burden of building background models independently for each domain, and are exploring methods that would enable us to avoid background-subtraction in the visible domain (RGB channels). In the future, we would also like to combine other mid-level fusion techniques (Section 2) into a background-subtraction framework to enable comparisons with our proposed algorithm. Further, as the approach is not limited to only extracting silhouettes of people, we will also evaluate the method for detecting other objects of interest (e.g., vehicles and animals).

Acknowledgments

This research was supported in part by the National Science Foundation under Grant No. 0236653. The authors thank Kyungnam Kim and Prof. Larry Davis for providing relevant code and the necessary parameters for the codebook technique for background-subtraction [24]. Lastly, a related, shorter version of the work presented here appeared in [12].

References

[1] T. Bakert, P. Losiewicz, Force aggregation via bayesian nodal analysis, in: Proceedings of Information Technology Conference, 1998, pp. 6–9.

[2] B. Bhanu, J. Han, Kinematic-based human motion analysis in infrared sequences, in: Proceedings on Workshop Applications of Computer Vision, 2002, pp. 208–212.

[3] B. Bhanu, R. Holben, Model-based segmentation of FLIR images, IEEE Transactions on Aerospace and Electronic Systems 26 (1) (1990) 2–11.

[4] P. Bogler, Shafer-dempster reasoning with applications to multisensor target identification systems, IEEE Transactions on System, Man, and Cybernetics 17 (1987) 968–977.

[5] D. Borghys, P. Verlinde, C. Perneel, 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 (1998) 477–484.

[6] P. Burt, R. Kolczynski, Enhanced image capture through fusion, in: Proceedings on Computer Vision and Pattern Recognition, 1993, pp. 173–182.

[7] A. Danker, A. Rosenfeld, Blob detection by relaxation, IEEE Transactions on Pattern Analysis and Machine Intelligence 3 (1) (1981) 79–92.

[8] S. Das, W. Krebs, Sensor fusion of multi-spectral imagery, Electronics Letters 36 (2000) 1115–1116.

[9] B. Dasarathy, Decision Fusion, IEEE Computer Society Press, 1994.

[10] J. Davis, V. Sharma, Robust background-subtraction for person detection in thermal imagery, in: IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2004.

[11] J. Davis, V. Sharma, Robust detection of people in thermal imagery, in: Proceedings on International Conference on Pattern Recognition, 2004, pp. 713–716.

[12] J. Davis, V. Sharma, Fusion-based background-subtraction using contour saliency, in: IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, 2005.

[13] J. Davis, V. Sharma, Background-subtraction in thermal imagery using contour saliency, International Journal of Computer Vision 71 (2) (2007) 161–181.

[14] R. Delaonoy, J. Verly, D. Dudgeon, Pixel-level fusion using interest images, Proceedings of the 4th National Symposium on Sensor Fusion, vol. 1, IRIA (ERIM), 1991, pp. 29–41.

[15] A. Elgammal, D. Harwood, L. Davis, Non-parametric model for background subtraction, in: Proceedings of the European Conference on Computer Vision, 2000, pp. 751–767.

[16] Francis Corbett et al., Fused atr algorithm development for ground to ground engagement, in: Proceedings of the 6th National Sensory Symposium, vol. 1, 1993, pp. 143–155.

[17] D.A. Fay et al., Fusion of multi-sensor imagery for night vision: color visualization, target learning and search, in: 3rd International Conference on Information Fusion, 2000, pp. TuD3–3–TuD3–10.

[18] R. Fletcher (Ed.), Practical Methods of Optimization, Wiley, New York, 1990.

[19] D. Gavrila, Pedestrian detection from a moving vehicle, in: Proceedings of the European Conference on Computer Vision, 2000, pp. 37–49.

[20] I. Haritaoglu, D. Harwood, L. Davis, W4: Who? When? Where? What? A real time system for detecting and tracking people, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 1998, pp. 222–227.

[21] M. Harville, A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 543–560.

[22] M. Hinman, Some computational approaches for situation assessment and impact assessment, in: Proceedings of the Fifth International Conference on Information Fusion, 2002, pp. 687–693.

[23] S. Iwasawa, K. Ebihara, J. Ohya, S. Morishima, Real-time estimation of human body posture from monocular thermal images, in: Proceedings on Computer Vision and Pattern Recognition, IEEE, 1997, pp. 15–20.

[24] K. Kim, T. Chalidabhongse, D. Harwood, L. Davis, Real-time foreground-background segmentation using codebook model, Real Time Imaging 11 (3) (2005) 167–256.

[25] L. Lazofson, T. Kuzma, Scene classification and segmentation using multispectral sensor fusion implemented with neural networks, in: Proceedings of the 6th National Sensor Symposium, vol. 1, 1993, pp. 135–142.

[26] C. Lemaréchal, R. Fjørtoft, Comments on geodesic saliency of waterhed contours and hierarchical segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (7) (1998) 762–763.

[27] J. Lewis, R. O'Callaghan, S. Nikolov, D. Bull, C. Cangarajah, Region-based image fusion using complex wavelets, in: International Conference on Information Fusion, 2004, pp. 555–562.

[28] H. Li, B.S. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graphical Model and Image Processing 57 (1995) 234–245.

[29] J. Lloyd, Thermal Imaging Systems, Plenum Press, New York, 1975.

[30] J. Lowrance, T. Garvey, T. Strat, A framework for evidential reasoning system, in: Proceedings of the Fifth National Conference on Artificial Intelligence, 1986, pp. 896–901.

[31] A. Mittal, N. Paragios, Motion-based background subtraction using adaptive kernel density estimation, in: Proceedings on Computer Vision and Pattern Recognition, 2004, pp. 302–309.

[32] A. Monnet, A. Mittal, N. Paragios, V. Ramesh, Background modeling and subtraction of dynamic scenes, in: Proceedings of the International Conference on Computer Vision, 2003, pp. 1305–1312.

[33] L. Najman, M. Schmitt, Geodesic saliency of watershed contours and hierarchical segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (12) (1996) 1163–1173.

[34] H. Nanda, L. Davis, Probabilistic template based pedestrian detection in infrared videos, in: Proceedings of the Intelligent Vehicles Symposium, IEEE, vol. 1, 2002, pp. 15–20.

[35] M. Pavel, J. Larimer, A. Ahumada, Sensor fusion for synthetic vision, in: AIAA Conference on Computing in Aerospace 8, 1991, pp. 164–173.

[36] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, S. Harp, Urban surveillance systems: from the laboratory to the commercial world, Proceedings of the IEEE 89 (10) (2001) 1478–1497.

[37] V. Petrovic, C. Xydeas, Gradient-based multiresolution image fusion, IEEE Transactions on Image Processing 13 (2) (2004) 228–237.

[38] G. Piella, A region-based multiresolution image fusion algorithm, in: Information Fusion, 2002, pp. 1557–1564.

[39] C. Ramac, M. Uner, P. Varshney, M. Alford, D. Ferris, Morphological filters and wavelet-based image fusion for concealed weapons detection, Proceedings of the SPIE 3376 (1998) 110–119.

[40] R. Raskar, A. Llie, J. Yu, Image fusion for context enhancement and video surrealism, in: Non-Photorealistic Animation and Rendering, ACM, 2004, pp. 85–94.

[41] C.V. Rijsbergen, Information Retrieval, second ed., Department of Computer Science, University of Glasgow, 1979.

[42] S. Russell, P. Norvig (Eds.), Artificial Intelligence: A Modern Approach, Prentice Hall, 2003.

[43] P. Scheunders, Multiscale edge representation applied to image fusion, in: Wavelet applications in signal and image processing VIII, 2000, pp. 894–901.

[44] Diego. A. Socolinsky, Lawrence. B. Wolff, A new visualization paradigm for multispectral imagery and data fusion, in: Proceedings on Computer Vision and Pattern Recognition, 1999, pp. 319–324.

[45] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings on Computer Vision and Pattern Recognition, IEEE, 1999, pp. 246–252.

[46] A. Toet, Hierarchical image fusion, Machine Vision and Applications 3 (1990) 1–11.

[47] K. Toyama, B. Brumitt, J. Krumm, B. Meyers, Wallflower: principals and practice of background maintenance, in: Proceedings of the International Conference on Computer Vision, 1999, pp. 49–54.
[48] P. Varshney, Distributed Detection and Data Fusion, Springer Verlag, 1996.
[49] L. Vincent, P. Soille, Watershed in digital spaces: an efficient algorithm based on immersion simulations, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (6) (1991) 583–598.
[50] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.
[51] F. Xu, K. Fujimura, Pedestrian detection and tracking with night vision, in: Proceedings of the Intelligence Vehicles Symposium, IEEE, 2002.
[52] A. Yilmaz, K. Shafique, M. Shah, Target tracking in airborne forward looking infrared imagery, Image and Vision Computing 21 (7) (2003) 623–635.
[53] M. Kokar, Z. Korona, Model-based fusion for multisensor target recognition, Proceedings of the SPIE 2755 (1996) 178–189.
[54] Z. Zhang, R. Blum. Region-based image fusion scheme for concealed weapon detection, in: Proceedings of the 31st Annual Conference on Information Sciences and Systems, 1997, pp. 168–173.
[55] J. Zhong, S. Sclaroff. Segmenting foreground objects from a dynamic, textured background via a robust kalman filter, in: Proceedings of the International Conference on Computer Vision, 2003, pp. 44–50.