Fusion-Based Background-Subtraction using Contour Saliency*

James W. Davis Vinay Sharma
Dept. of Computer Science and Engineering
Ohio State University
Columbus OH 43210 USA
{jwdavis,sharmav}@cse.ohio-state.edu

Abstract

We present a new contour-based background-subtraction technique using 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 combined. An A path-constrained search along watershed boundaries is used to complete and close any broken contour segments. Lastly, the contour image is flood-filled to produce silhouettes. Results of our approach are presented and compared against manually segmented data.*

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). To meet this ideal, we present a new background-subtraction technique for object detection that relies on two complementary bands of the electromagnetic spectrum, long-wave infrared (thermal) and visible light.

Thermal (FLIR) video 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 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, 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 very hot or cold objects. 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 object (person) is much hotter than the surrounding environment. Though this is common in 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 contour-based background-subtraction algorithm using both visible and thermal imagery. The approach is 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 [3, 2] for object detection in thermal imagery.

The proposed technique assumes that the two image streams, thermal and visible, are co-registered. Using a standard background-subtraction technique, we first identify regions-of-interest (ROIs) in the thermal domain. ROIs in the visible domain are then obtained by performing color- and intensity-based background-subtraction within the regions identified in the thermal domain. Within each image region (thermal and visible treated independently), the input and background gradient information are combined as to highlight only the boundaries of the foreground object. The boundaries are then thinned and thresholded to form binary contour fragments. Contour fragments belonging to corresponding regions in the thermal and visible domains are then fused using the combined input gradient informa-

*Appears in *IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, San Diego, CA, June 20, 2005.

tion from both sensors. An A* search algorithm constrained to a local watershed segmentation is then used to complete and close any contour fragments. Finally, the contours are flood-filled to make silhouettes (for later use in shape-based activity analysis).

We demonstrate the approach using a single set of parameters/thresholds across different thermal and color video sequences recorded from two different locations on a university campus. Based on a set of manually segmented images, we also quantitatively compare the results obtained by fusing visible and thermal information against using either of the two sensors alone.

2. Related Work

Several object detection strategies have been proposed for the visible domain. Approaches focussing on person detection that do not use background-subtraction methodologies, include the direct use of wavelets [9], coarse-to-fine edge matching [5], and motion differencing [16]. Most of the other object detection methods employ some form of background-subtraction using a single Gaussian background model [17] or a multimodal Gaussian formulation [13]. Other approaches include the three-stage Wallflower approach [15], a two-stage color and gradient technique [6], and a Markov chain Monte Carlo approach [19].

Recently, person detection and tracking using thermal imagery has been explored [1, 8, 18], but these approaches rely on the highly limiting assumption that the person region always has a much brighter (hotter) appearance than the background.

Image fusion techniques have had a long history in vision. Gradient-based techniques include defining first-order contrasts in high dimensions [12] and examining gradients at multiple resolutions [11]. Several region-based multiresolution algorithms have been proposed such as the pyramid approaches of [14, 10] and the wavelet-based approach of [7]. Other biologically motivated techniques [4] have also been proposed. Most of these fusion techniques aim at enhancing the information content of the scene, to ease and enhance human visual analysis. However, the method we propose is designed specifically to enhance the capabilities of an automatic vision-based detection system.

This paper extends our prior work presented in [3, 2], and provides a framework for effectively combining information from thermal and visible imagery.

3. Initial Region Detection

Since the algorithm requires registered imagery from the two sensors, we initialize the system by manually selecting 12 corresponding feature points from a pair of thermal and visible images. A homography matrix created from these points is used to register the thermal and visible images (other techniques could also be applied).

We begin the process by identifying localized regions-of-interest (ROIs) in both domains (thermal and visible). The background in the thermal domain tends to be more stable over time (changing slowly with environmental variations), and standard background-subtraction generally produces regions that encompass the entire foreground object (and surrounding halo). Therefore we use the ROIs in the thermal imagery to cue further processing in both the thermal and visible domains (ROIs in the visible domain will be extracted within the selected thermal ROIs).

To construct 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 frames) for the thermal images and for the 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

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

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

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

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

with a standard deviation $\widehat{\sigma} = 5$. The farther $I_i(x,y)$ is from $I_{med}(x,y)$, the smaller its contribution.

In the thermal domain, once the statistical background model has been constructed, we obtain the foreground pixels for any new input image I using the squared Mahalanobis distance

$$D^T(x,y) = \begin{cases} 1 & \frac{(I(x,y)-\mu(x,y))^2}{\sigma(x,y)^2} > 10^2 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where the superscript T denotes the thermal domain.

To extract the thermal ROIs, 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.

The image region in the visible domain corresponding to a thermal ROI might contain object shadows. Hence color and intensity information are exploited *within* each

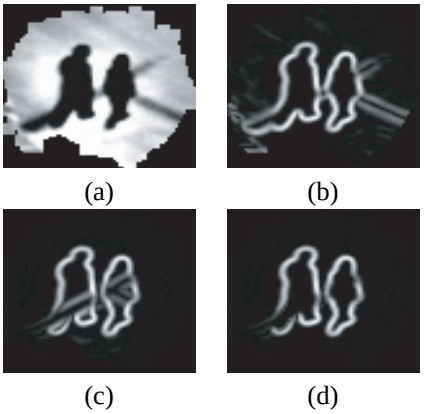


Figure 1: Contour saliency. (a) Thermal ROI. (b) Input gradient magnitudes. (c) Input-background gradient-difference magnitudes. (d) CSM.

(thermal) ROI to identify true object regions in the visible domain (without shadows). The intensity component is used to identify those regions (D^{Int}) in an input visible image that are brighter than the background, based on the squared Mahalanobis distance, with a threshold of 10 SD. The mean/variance model for the intensity component is computed using Eqns. 1 and 2. The normalized RGB components are used to detect regions (D^{Col}) in the input image that are different in color from the background, based again on the squared Mahalanobis distance with a threshold of 4 SD. A mean/covariance model of the normalized color-space is computed directly from the initial set of N visible images (without the weights in Eqn. 3). The D^T obtained from the thermal domain is then used as a mask over the intensity/color background-subtraction results.

$$D^V = (D^{Int} \cup D^{Col}) \cap D^T \tag{5}$$

where the superscript V denotes the visible domain. Similar to the thermal domain, a 5×5 dilation operator is applied to D^V. Having obtained D^V, we make use of only the intensity components of both the thermal and visible domains for the following contour detection process.

4. Contour Detection

We next 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) [2], 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

$$\text{CSM} = \min\left(\frac{\|\langle I_x, I_y\rangle\|}{Max}, \frac{\|\langle (I_x - BG_x), (I_y - BG_y)\rangle\|}{Max}\right) \tag{6}$$

where the normalization factors 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 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 ROI in Fig. 1. The gradients were calculated using 7×7 Gaussian derivative masks.

4.1. 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 we use the non-maximum suppression result of the *input* gradients as a thinning mask for the CSM.

4.2. Thresholding

After thinning, we threshold the tCSM to select the most salient contour segments. The approach is motivated by object properties in the thermal domain, where foreground object pixels are either unimodal or multimodal, but the technique has also shown to be applicable to visible imagery.

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 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 hypothesis

is that an optimally thresholded tCSM should contain contours of relatively high average length that also "cover" the ROI sufficiently well.

We compute the ACL for a tCSM 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 ROI perimeter pixels to the closest pixel in the thresholded tCSM as a measure of the coverage (C). Thus the quality of a thresholded image B_i is evaluated using

$$Q(B_i) = (1-\alpha)\cdot\left(\frac{ACL(B_i)}{\max(ACL(B_2), ACL(B_3))}\right) + \alpha\cdot\left(1 - \frac{C(B_i)}{\max(C(B_2), C(B_3))}\right) \quad (7)$$

The binary image (B_2, B_3) that maximizes Q is chosen as the best thresholded result.

The weighting factor α determines the influence of each factor in Q. Empirically, we found that the weight α should be a function of the ratio of the two ACLs

$$r = \frac{\min(ACL(B_2), ACL(B_3))}{\max(ACL(B_2), ACL(B_3))} \quad (8)$$

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

$$\alpha = \frac{1}{1 + e^{-\beta\cdot(r-0.5)}} \quad (9)$$

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

In Fig. 2 (top row) 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 ACL was the dominating factor in the quality evaluation. In Fig. 2 (bottom row), we show a thermal ROI with *multimodal* person pixels and its binary images. 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 almost identical.

5. 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. We first

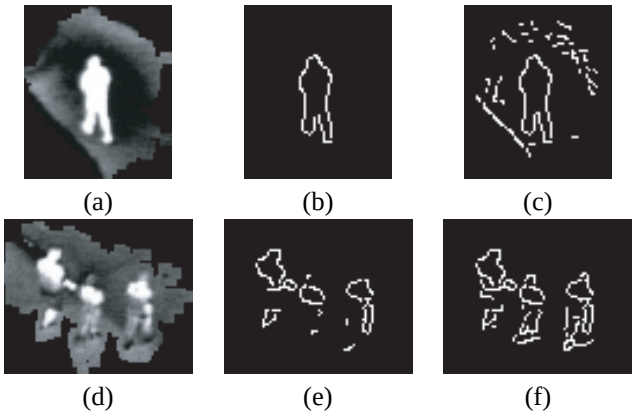


Figure 2: Contour selection. Top row: (a) Unimodal thermal ROI. (b) B_2 (selected). (c) B_3. Bottom row: (d) Multimodal thermal ROI. (e) B_2. (f) B_3 (selected).

combine the information from the two sensors by performing a simple union of their individual contributions using

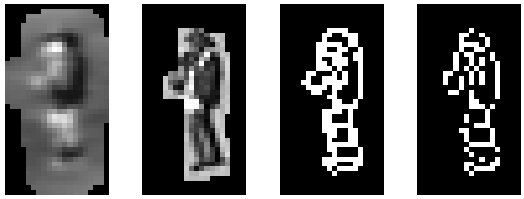
$$\text{tCSM}_b = \text{tCSM}_b^T \cup \text{tCSM}_b^V \quad (10)$$

where the subscript b denotes that the tCSM is binary. Since our thermal and visible images were registered based on a limited homography matrix, there is no guarantee that contour fragments belonging to the same edge (extracted individually in the two domains) will exactly coincide with each other. Thus the tCSM$_b$ needs to be further aligned such that only those contour fragments that correspond to gradient maxima across both domains are preserved.

To achieve this, we first create a combined input gradient map from the foreground gradients of each domain. Gradient direction and magnitude information at a pixel in tCSM$_b$ is selected from either the thermal or the visible domain depending on it being present in tCSM$_b^T$ or tCSM$_b^V$. If present in both, the gradient information at that pixel can be taken from either domain. Since we now have orientation and magnitude at every contour pixel in the tCSM$_b$, we apply a local non-maximum suppression algorithm to perform a second thinning to better align the 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. 3(a) and Fig. 3(b) we show corresponding ROIs in the thermal and visible domains. Figures 3(c) and (d) show the tCSM$_b$ before and after alignment, respectively.

6. Contour Completion and Closing

While contour information from the two channels are often complementary, the contour fragments in the combined tCSM$_b$ are still mostly broken 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

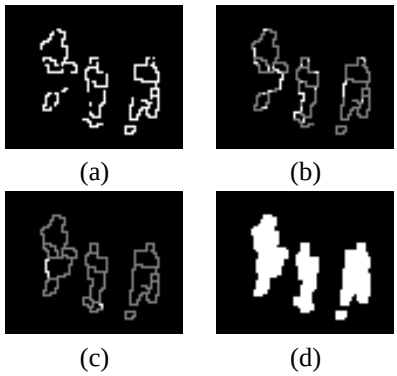


(a) (b) (c) (d)

Figure 3: Fused binary contours. (a) Thermal ROI. (b) Visible/intensity ROI. (c) Fused binary contours before alignment. (d) Fused binary contours after alignment.

we can apply the flood-fill operation to create silhouettes. We use the two-stage method first suggested in [3, 2] to perform this task. We provide here only a brief description of the basic algorithm.

The first stage of the approach completes any contour gaps by using the A* search algorithm from each contour endpoint to find another contour pixel. The watershed lines of the input gradient image are used to limit the search space to only meaningful paths. To eliminate small stray contour fragments present in the tCSM_b that may harm the completion/closing processes, we extend the approach by first obtaining a coarser segmentation of the ROI using a basin-merging algorithm on the watershed partition. We employ 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). Based on the merged watershed segmentation, the tCSM_b is partitioned into distinct segments that divide pairs of adjacent basins. A contour segment is removed if its length is less than 50% of the length of the watershed border separating the two basins. After the completion process, the second stage (closing) ensures that every contour in the image is part of a closed loop (using a joining approach similar to the prior completion method). Lastly, the result is flood-filled to produce the silhouettes. We show step-by-step results for a fragmented contour ROI in Fig. 4.



(a) (b) (c) (d)

Figure 4: 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.

7. Experiments

To examine our contour-based fusion approach, we tested the 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. The number of frames in each sequence are Sequence-1:2107, Sequence-2:1201, Sequence-3:3399, Sequence-4:3011, Sequence-5:4061, and Sequence-6:3303. 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. Example images from this dataset, one from each sequence, are shown in the top rows of Fig. 6. The sequences were recorded on the Ohio State University campus, 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. To incorporate variations in the thermal domain, the gain/level settings on the thermal camera were varied across the sequences. The images of Sequences 4, 5, and 6, were captured on a cloudy day, with fairly constant illumination and soft/diffuse shadows.

To demonstrate the generality and applicability of our approach, we extracted silhouettes from each pair of sequences with the proposed method using the **same parameter/threshold settings for all sequences**. For each silhouette, we computed a contrast value in each of the separate domains (thermal, visible) as the ratio of the maximum input-background intensity difference within the silhouette region to the full intensity range of the background image. The final contrast value for each silhouette was the larger of these two ratios. A final user-selected threshold can be used to remove any minimal-contrast (noise) regions.

To quantitatively measure the performance of the detection results, we obtained a manual segmentation of the person regions in 30 image-pairs from our dataset (5 image-pairs spanning each of the 6 sequences). For each of the 30 image-pairs, 3 people hand-segmented the person regions, 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. Six of the median silhouette images

are shown in Fig. 6.

Using the manually segmented images, we compared results of the proposed approach under three different input scenarios: both thermal and visible imagery (T-V), only thermal imagery (OT), and only visible imagery (OV). The OT scenario would be similar to nighttime surveillance when the visible sensor would be ineffective. The background models for the visible and thermal images were computed using the technique described in Sect. 3 with N set to the length of the sequence being processed.

To quantitatively compare the algorithm results across the three scenarios with the manually segmented images, we examined *Sensitivity* and *Positive Predictive Value* (PPV) measurements. Sensitivity refers to the fraction of object/person pixels that are correctly detected by the algorithm, while PPV represents the fraction of detections that are in fact object/person pixels. For each scenario, only the final silhouette contrast threshold was adjusted (the remaining parameters were fixed) over a large range and the threshold yielding the largest sum of the Sensitivity and PPV over all of the 30 test images was selected. The results showed quite reasonable silhouettes throughout the sequences. We show silhouette results obtained by the proposed algorithm under each scenario on six representative images (from the set of 30) in Fig. 6.

The Sensitivity and PPV results for the 30 images are shown in Table 1. Both OT and OV have high PPVs. The high PPVs obtained in OT demonstrate the efficacy of the contour-based approach, which enables detection of person information within large thermal halo regions. For OV, the high PPVs suggest that most shadow regions were successfully eliminated. However, incorrect shadow removal and frequent illumination changes are the main causes of the low Sensitivity of OV. Urban environments often have poor color information essential for accurate shadow detection/removal, and hence some person regions darker than the background can be eliminated as shadows. Other more computationally expensive approaches could however be employed.

The best Sensitivity rate of OV is for Sequence 6, when conditions were overcast and no prominent shadows were present. Some image regions (taken from the set of 30) of the visible domain are shown in the top row of Fig. 5. Figure 5(a) shows a challenging case where the person appears dark and is in shadow, Fig. 5(b) shows a person casting a prominent shadow, and Fig. 5(c) shows another example of a person walking through a dimly lit region and casting a shadow.

Thermal images corresponding to the color images of Fig. 5 (a), (b), and (c) are shown in the second row of the figure. In Fig. 5(a) the person region appears white hot, possibly due to the lower temperatures in the shade. Figure 5(b) shows a person region with different thermal characteristics and Fig. 5(c) shows an example of a person region close to the thermal cross-over.

(a) (b) (c)

Figure 5: Three example person regions in visible (top row) and thermal imagery (middle row), and results after contour fusing (bottom row).

The bottom row of Fig. 5 shows the results of the proposed approach, before contour completion. The contour images are coded in gray and white, corresponding to higher contrasts in the thermal and visible domains, respectively. While the thermal domain dominates in Fig. 5(a), the visible domain provides stronger information for the person's legs in Fig. 5(b). In Fig. 5(c) the visible domain is less dominant due to misclassification of person regions as shadows.

The Sensitivity rates of OT are better than OV. This is to be expected since it is unlikely that exposed parts of the human body are at thermal cross-over. However, clothing that insulates body heat (e.g., thick winter jackets), can quickly attain thermal equilibrium with the surroundings, making person regions harder to discern in the thermal domain.

The Sensitivity rates for T-V are the best for all of the sequences, while the PPV is not significantly compromised. This shows that the proposed algorithm is able to exploit the complementary nature of the sensors. The average results in the final column of Table 1 demonstrate that the best overall combined score of S and PPV is obtained when both visible and thermal images are used.

8. 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 [3]. 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 used 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 was then used within these regions to extract the corresponding regions-of-interest (without shadows) in the visible domain. The input and background gradient information within each region were then combined into a Contour Saliency Map (CSM). The CSM was thinned using a non-maximum suppression mask of the individual input gradients. The most salient contours were then selected using a thresholding strategy based on competitive clustering. The binary contours from corresponding regions of the thermal and visible image were then combined, and thinned using the input gradient information from both sensors. Any broken contour fragments were completed and closed using a watershed-constrained A* search strategy. Lastly, the contours were flood-filled to produce silhouettes.

Experiments were conducted with six challenging thermal/color video sequences recorded from different locations and at different times-of-day. A manually segmented subset of 30 images was used to compare the results of our algorithm with thermal-only, visible-only, and combined thermal-visible inputs. Our method, using a single set of parameters/thresholds, showed promising results. The quantitative results using Sensitivity and Positive Predictive Value demonstrated the enhanced performance obtained by fusing visible and thermal imagery with our approach.

To further improve our results, we plan to include motion information into the saliency map, and employ shaped-based models for better figure completion and tracking. Furthermore, we will incorporate an adaptive background model to test our algorithm over longer durations. As the approach is not limited to only extracting silhouettes of people, we will also examine the method for detecting other objects of interest (e.g., vehicles and animals).

9. Acknowledgements

This research was supported in part by the National Science Foundation under grants No. 0236653 and 0428249, and the Secure Knowledge Management Program, Air Force Research Laboratory (Information Directorate, WPAFB, OH).

References

[1] B. Bhanu and J. Han. Kinematic-based human motion analysis in infrared sequences. In *Proc. Wkshp. Applications of Comp. Vis.*, pages 208–212, 2002.

[2] J. Davis and V. Sharma. Robust background-subtraction for person detection in thermal imagery. In *IEEE Int. Wkshp. on Object Tracking and Classification Beyond the Visible Spectrum*, 2004.

[3] J. Davis and V. Sharma. Robust detection of people in thermal imagery. In *Proc. Int. Conf. Pat. Rec.*, pages 713–716, 2004.

[4] D. A. Fay et al. Fusion of multi-sensor imagery for night vision: Color visualization, target learning and search. In *Int. Conf. on Info. Fusion*, 2000.

[5] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. European Conf. Comp. Vis.*, pages 37–49, 2000.

[6] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Wkshp. on Motion and Video Computing*, pages 22–27. IEEE, 2002.

[7] H. Li, B. Manjunath, and S. Mitra. Multisensor image fusion using the wavelet transform. In *Graphical Models and Image Processing*, volume 57, pages 234–245, 1995.

[8] H. Nanda and L. Davis. Probabilistic template based pedestrian detection in infrared videos. In *Proc. Intell. Vehicles Symp.* IEEE, 2002.

[9] M. Oren, C. Papageorgiour, P. Sinha, E. Osuma, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. Comp. Vis. and Pattern Rec.*, pages 193–199. IEEE, 1997.

[10] M. Pavel, J. Larimer, and A. Ahumada. Sensor fusion for synthetic vision. In *Conf. on Computing in Aerospace*. AIAA, 1991.

[11] P. Scheunders. Multiscale edge representation applied to image fusion. In *Wavelet Applications in Signal and Image Processing VIII*, pages 894–901, 2000.

[12] D. Socolinsky and L. Wolff. A new visualization paradigm for multispectral imagery and data fusion. In *Proc. Comp. Vis. and Pattern Rec.*, pages 319–324, 1999.

[13] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Comp. Vis. and Pattern Rec.*, pages 246–252. IEEE, 1999.

[14] A. Toet. Heirarchical image fusion. *Machine Vision and Applications*, 3:1–11, 1990.

[15] K. Toyama, B. Brumitt, J. Krumm, and B. Meyers. Wallflower: principals and practice of background maintenance. In *Proc. Int. Conf. Comp. Vis.*, pages 49–54, 1999.

[16] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. Int. Conf. Comp. Vis.*, pages 734–741, 2003.

[17] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 19(7):780–785, 1997.

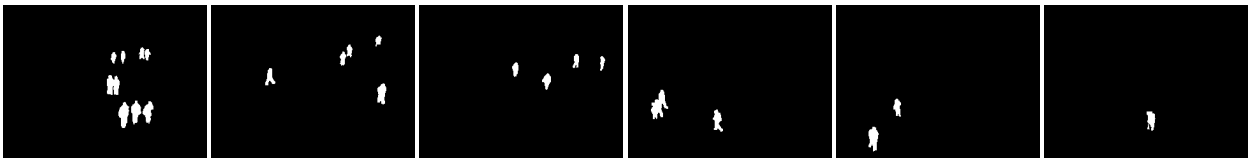
Method		Seq-1	Seq-2	Seq-3	Seq-4	Seq-5	Seq-6	Average
Thermal and Visible	S	0.773	0.772	0.699	0.695	0.852	0.817	0.764
	PPV	0.886	0.835	0.898	0.939	0.933	0.915	0.887
Thermal only	S	0.629	0.684	0.578	0.685	0.788	0.722	0.658
	PPV	0.943	0.922	0.919	0.958	0.940	0.913	0.933
Visible only	S	0.527	0.246	0.439	0.280	0.441	0.725	0.435
	PPV	0.855	0.928	0.860	0.932	0.969	0.956	0.889

Table 1: Comparison of detection results using Sensitivity (S) and Positive Predictive Value (PPV).

[18] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. In *Proc. Intell. Vehicles Symp.* IEEE, 2002.

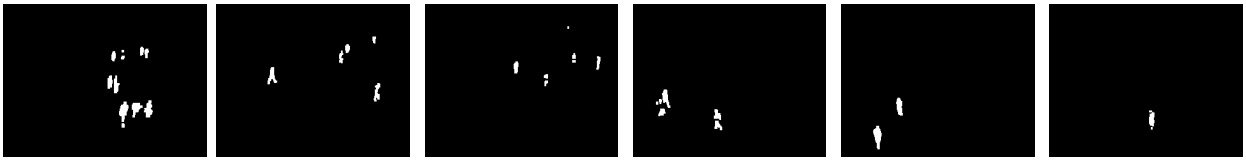
[19] T. Zhao and R. Nevatia. Stochastic human segmentation from a static camera. In *Wkshp. on Motion and Video Computing*, pages 9–14. IEEE, 2002.

Examples of registered thermal and visible image pairs



Manually segmented silhouettes

Proposed method, using both thermal and visible domains



Proposed method, using only thermal domain

Proposed method, using only visible domain

Figure 6: Visual comparison of detection results of the proposed approach across different images and scenarios.