International Journal of Computer Vision 71(2), 161–181, 2007

DOI: 10.1007/s11263-006-4121-7

Background-Subtraction in Thermal Imagery Using Contour Saliency

JAMES W. DAVIS AND VINAY SHARMA
Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210, USA
jwdavis@cse.ohio-state.edu
sharmav@cse.ohio-state.edu

Received April 18, 2005; Revised August 1, 2005; Accepted August 10, 2005

First online version published in June, 2006

Abstract. We present a new contour-based background-subtraction technique to extract foreground objects in widely varying thermal imagery. Statistical background-subtraction is first used to identify local regions-of-interest. Within each region, input and background gradient information are combined to form a Contour Saliency Map. After thinning, 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 for several difficult thermal sequences and compared to alternate approaches. We quantify the results using manually segmented thermal imagery to demonstrate the robustness of the approach.

Keywords: background subtraction, thermal imagery, infrared, FLIR, contour saliency map, CSM, video surveillance and monitoring, person detection

1. Introduction

We present a new background-subtraction technique to robustly extract foreground objects in thermal video under different environmental conditions. Thermal (FLIR) video cameras detect relative differences in the amount of thermal energy emitted/reflected from objects in the scene. As long as the thermal properties of a foreground object are slightly different (higher or lower) from the background radiation, the corresponding region in a thermal image appears at a contrast from the environment. Therefore thermal cameras can be equally applicable to both day and night scenarios, making them a prime candidate for a persistent (24-7) video system for surveillance and monitoring. Thermal cameras have been traditionally used by the military for tasks such as long-range detection of enemy vehicles and Automatic Target Recognition (ATR). In recent years, thermal cameras have become increasingly employed for other applications, including industrial inspection, surveillance, and law enforcement.

The use of thermal imagery alleviates several classic computer vision problems such as the presence of shadows (which appear in the thermal domain only when an object is stationary long enough for the shadow to cool the background), lack of nighttime visibility, and sudden illumination changes. However, thermal imagery has its own unique challenges, including a lower signal-to-noise ratio, uncalibrated white-black polarity changes, and the "halo effect" that appears around very hot or cold objects in imagery produced by common ferroelectric BST sensors. In Fig. 1 we show outdoor surveillance images of the same scene captured with a thermal camera, but taken on different days (morning and afternoon). The thermal properties of the people and background are quite different, including the change from bright (hot) people to dark (cool) people in relation to the background. For such thermal imagery to be used reliably in automatic urban

Figure 1. Thermal images showing large variation in polarity and intensity.

surveillance, these image variations must be properly addressed.

Most of the previous strategies for object/person detection in thermal imagery employ "hot-spot" algorithms, relying on the 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 across different seasons of the year (as convincingly shown in Fig. 1). As we will show, standard background-subtraction, image-differencing, and hot-spot techniques are by themselves ineffective at extracting the precise locations and shapes of people in such diverse imagery.

A prominent characteristic of thermal imagery produced from uncalibrated ferroelectric BST (chopper) sensors is the presence of halos around objects having a high thermal contrast with the background (Hoist, 2000). The halos have the opposite polarity (light/dark) of the objects they surround. The strength and size of the halos depend on factors such as the actual temperature differential between the object and the surrounding environment and the contrast/gain setting on the camera. While this poses the most significant challenge to existing (and popular) background-subtraction methodologies, we will show that our method in turn capitalizes on this unavoidable artifact to improve performance.

It should be noted here that microbolometer thermal sensors, however, do not produce the haloing effect. In spite of this advantage, several signal-based factors make traditional ferroelectric BST sensors more favorable (Kummer, 2003). Ferroelectric BST sensors, being AC-coupled, are better equipped to handle detector-induced steady-state noise which can have a significant negative impact on image quality. Also, they are capable of resolving greater temperature variations in a scene than the DC-coupled microbolometers. Further, microbolometers require to recalibrate the scene at random intervals to minimize spatial noise. This can be a serious drawback to vision-based systems as the video output freezes momentarily during each recalibration. Lastly, commercially available microbolometers are typically half the resolution of ferroelectric BST sensors (with higher resolutions being considerably more expensive). More detailed comparisons of the two sensors can be found in Kummer (2003), Pandya and Anda (2004). Due to the aforementioned issues, and that ferroelectric BST sensors are still in wide use by military and law enforcement agencies, algorithms targeted for use in the thermal domain need to be robust to the image characteristics of ferroelectric BST sensors.

The image characteristics of thermal halos produced by ferroelectric BST sensors can be examined using the intensity profile of a row of pixels sliced through an image region containing a halo. In Fig. 2(a), we show image regions containing people, one recorded in Summer (top), and the other recorded in Winter (bottom). The images were collected from different thermal cameras and at different ranges. Notice the polarity changes (white/black) for the people in the two images. In Fig. 2(b), we show the corresponding background regions without the people. The row of pixels to be examined is marked with a solid (Fig. 2(a)) and dotted (Fig. 2(b)) line and two boundary points (left and right) of the people have been marked with a circle. The plots in Fig. 2(c) show the intensity profiles for the corresponding input and background regions along the slice. Comparison of the input intensity profiles with those of the background makes it clear that the input region slice is brighter/darker *around* the people due to the halo, thus making the outer boundary contrast of the people stronger.

Two key observations can be made about thermal halos based on these plots: (1) thermal halos fade smoothly/slowly into the image, and (2) stronger halos

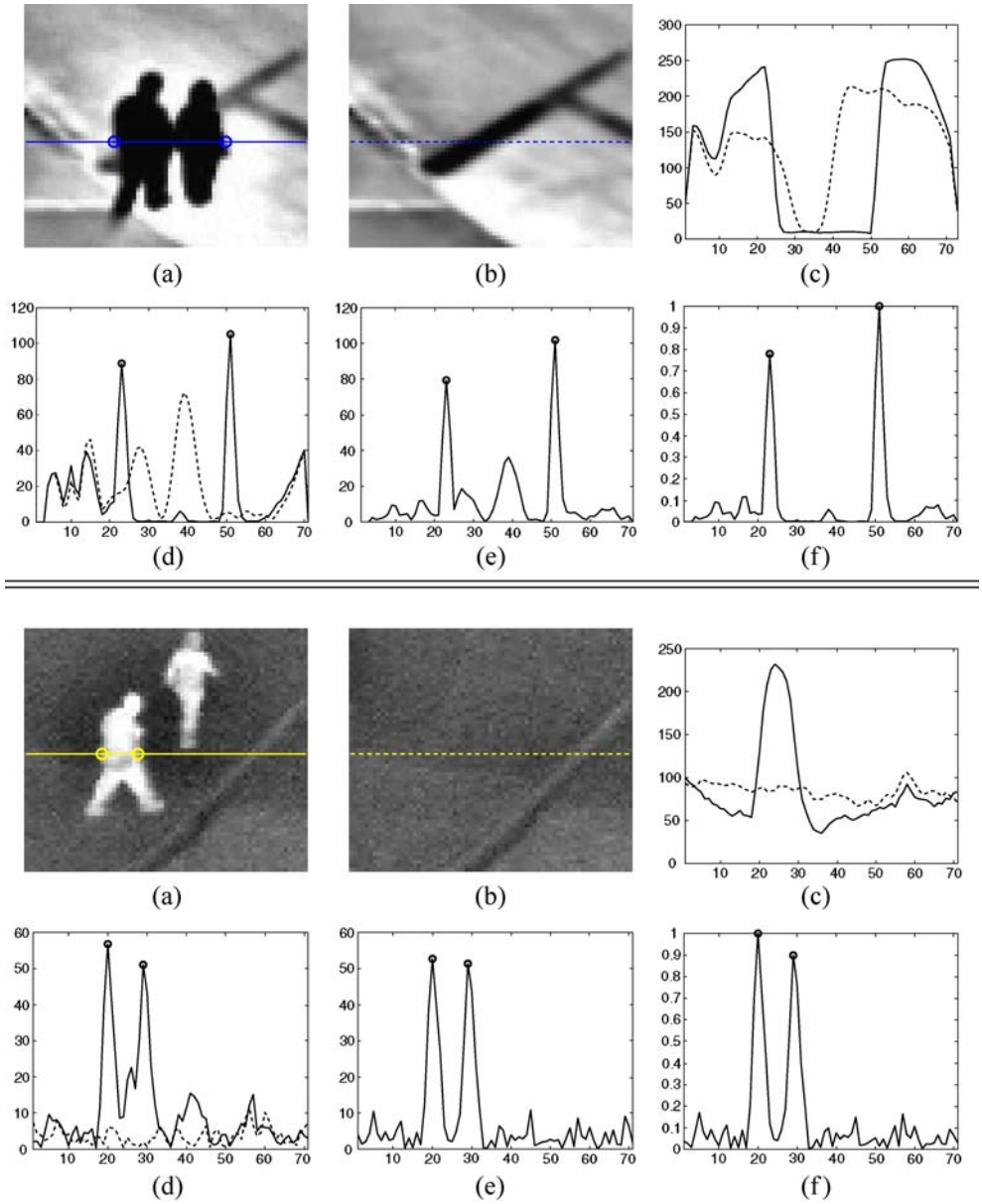

Figure 2. Characteristics of thermal halos. (a) Input region. (b) Background region. (c) Intensity profiles. (d) Gradient magnitude profiles. (e) Input-background gradient-difference magnitude profiles. (f) CSM profiles.

cause the gradient magnitude around the boundary of the object (within the halo) to become stronger. The input and background intensity profiles also predict the inability of standard background-subtraction methods to extract precise silhouette shapes from such thermal imagery (as the halo would be statistically-different from the background). Additionally, the image region shown in the top row of Fig. 2(a) makes it clear that hot-spot methods are not universally applicable, as the thermal intensity of the people is much *lower* (darker) than the background. Indeed, people do not always appear uniformly brighter or darker than the environment, and many times appear multimodal. Based on these observations, we exploit this normally detrimental halo

artifact of the sensor by focusing on the salient gradient information which in fact becomes more pronounced in the presence of halos. Furthermore, the gradient magnitude is invariant to white/black polarity differences.

Our proposed approach first uses a standard background-subtraction technique to identify local regions-of-interest (including the foreground objects and halos). The input and background gradient information within each region are then combined as to highlight only the foreground object boundary. This boundary is then thinned and thresholded to form contour fragments. An A* search algorithm constrained to a local watershed segmentation of the region is then used to complete and close any contour fragments. Finally, the contours are flood-filled to make silhouettes. As the approach relies on gradients rather than raw intensity, the method is more stable and robust across very different environmental conditions, including different halo strengths and intensity polarity switches. We demonstrate the approach across six very different thermal video sequences recorded from two different thermal cameras. We also quantitatively compare our results with three other common approaches using a set of manually segmented thermal images.

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. Though our method is a *foreground detection* approach and not an *object recognition* technique, the results from our approach could also be potentially used in a recognition scheme using silhouettes or contours (as in Davis and Keck (2005)). Furthermore, the proposed approach is not limited to ferroelectric BST sensors. Though the method is designed to handle thermal halos, it can be equally applied to non-halo imagery from other video sensors such as microbolometers or color CCDs. The effectiveness of the algorithm on color imagery is demonstrated in Section 4.2. We show that the method is capable of removing soft shadows around foreground objects in color imagery, as soft/diffuse shadows have a response similar to thermal halos in gradient-space.

The remainder of this paper is described as follows. We begin with a review of related work (Section 2). We then describe the main components of the proposed method (Section 3). Next we present experimental and comparative results (Section 4). Finally, we conclude with a summary of the research and discuss future work (Section 5).

2. Related Work

There exist several approaches to the problem of object detection/extraction in video. These methods can be grouped into two broad classes: template-based detection and background-subtraction. We discuss related work for each of these frameworks in both color and thermal imagery.

Regarding template-based methods, of particular interest are those designed for detecting people in video. The use of a color and texture invariant wavelet template, followed by a Support Vector Machine classifier, was proposed in Oren et al. (1997). In Gavrila (2000), a shape-based method using hierarchical templates and coarse-to-fine edge matching was used. Motion cues have also been exploited to identify pedestrians in video. In Cutler and Davis (1999), the periodicity and symmetry of humans walking were used to detect and classify people. In Viola et al. (2003), image differencing was combined with intensity information to create an AdaBoosted classifier to detect pedestrians. Two simple properties (dispersedness, area) were used by Lipton et al. (1998) to classify regions selected from image differencing as people or vehicles. These template methods must be trained with examples of all possible targets to be detected, and also do not extract the precise shape of the detected person (other than the approach of Gavrila (2000)). Generating accurate silhouettes enables the use of shape-based techniques by higher-level vision modules for tasks such as pose and activity recognition.

The popular non-template framework, not limited to the detection of particular shapes (e.g., people), is background-subtraction. 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 (Wren et al., 1997; Haritaoglu et al., 1998), 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 Stauffer and Grimson (1999) to better model the complex background processes of each pixel. The Mixture of Gaussians approach was also examined in Harville (2002). Other background-subtraction methods based on nonparametric statistical modeling of the pixel process have also been proposed. In Elgammal et al. (2000), kernel density estimation was used to obtain the pixel intensity

distributions. A variable-bandwidth kernel density estimator was proposed in Mittal and Paragios (2004). Time series analysis of input video is another technique used to create dynamic background models. Kalman filters were used in Zhong and Sclaroff (2003), and an auto-regressive model was used in Monnet at al. (2003). Weiner filters were employed in the three-stage (pixel/region/frame) Wallflower approach of Toyama et al. (1999).

The presence of halos in thermal imagery will severely impair the performance of each of the above methods as the halo artifact around foreground objects is typically much different than the expected background. Since the halo surrounding an object would also be detected as part of the foreground, the result would not provide an accurate localization of the object silhouette, when ironically the object shape is most easily distinguishable to human observers *because* of the halo. Some of the above methods (e.g., Haritaoglu et al., 1998; Gavrila, 2000) have been tested with thermal imagery, but the limited nature of the examples examined does not provide a comprehensive evaluation.

Many other algorithms focussing specifically on the thermal domain have been explored. The unifying assumption in most of these methods is the belief that the objects of interest are warmer than their surroundings, and hence appear brighter, as "hot-spots", in thermal imagery. In Iwasawa et al. (1997), Bhanu and Han (2002), a thresholded thermal image forms the first stage of processing after which methods for pose estimation and gait analysis are explored. In Nanda and Davis (2002), a simple intensity threshold is employed and followed by a probabilistic template. A similar approach using Support Vector Machines is reported in Xu and Fujimura (2002). The use of the strong hot-spot assumption can also be found in other work related to object detection and tracking in thermal imagery (Bhanu and Holben, 1990; Danker and Rosenfeld, 1981; Yilmaz et al., 2003). The underlying and limiting hot-spot assumption will be violated in imagery recorded at different environmental temperatures and in most urban environments (see Fig. 1).

The approach presented in this paper is based on our prior work in Davis and Sharma (2004a,b). The motivation of that work was to remove the invalid hot-spot assumption of the other approaches and to deal directly with the halo and polarity artifacts present in ferroelectric BST thermal imagery. The approach used background-subtraction and employed the Contour Saliency Map (CSM) representation to combine gradient information from the input and background images to focus on the strong object contours within halo regions and form silhouettes. In this paper, we extend the approach and quantitatively examine the method using several difficult variations of thermal imagery. The two-stage color and gradient technique of (Javed et al., 2002) is the most related to our approach in that it uses color information to initially detect foreground regions and then employs input-background gradient-differences (edges) to validate these regions. However thermal halos *surrounding* the object (not separate from the object) would cause difficulties with their approach.

Our approach employs several new methods to address the problem domain. The CSM representation (Section 3.2) is an original contribution to background-subtraction. A recent template-based approach to detect pedestrians in thermal imagery Davis and Keck (2005) was robust mainly due to the use of the CSM representation in the initial detection stage. The proposed "competitive clustering" method (Section 3.2) used in our approach to perform the selection of the most salient contours from the CSM is novel and based on common characteristics of thermal imagery. Additionally, our application of the watershed transform for robust contour completion (Section 3.3) is a novel use of the technique.

3. Method

3.1. Initial Region Detection

We begin the process by identifying the separate regions in an image that are likely to contain foreground objects (e.g., people). These regions-of-interest (ROIs), that include the foreground objects and the surrounding thermal halo, are obtained using a standard intensity-based statistical background-subtraction method with a background mean and variance model at each pixel. Other statistical approaches to model the background could also be used, such as Mixture of Gaussians (Stauffer and Grimson, 1999), but will not be sufficient to address the halo artifact.

In order to build a proper mean/variance background model, it is often difficult or infeasible to get a sufficiently long video clip without any foreground objects appearing in the scene. To overcome this problem, we first capture N images at a frame rate considerably lower (approximately 3 Hz) than that of the actual

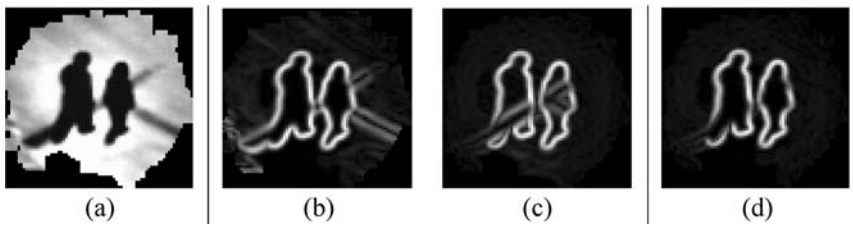

Figure 3. Contour saliency. (a) ROI. (b) Input gradient magnitudes. (c) Input-background gradient-difference magnitudes. (d) CSM.

frame rate of the camera, and create a median image (I_{med} from the N frames). As foreground objects could be present in the images, the statistical background model for each pixel is created by computing *weighted* means and variances from the N sampled values

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

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

where the weights $w_i(x, y)$ for each pixel location are used to minimize the effect of outliers (values far from the median $I_{\text{med}}(x, y)$), which in our case are the pixels belonging to the foreground objects and halos. The weights are computed from a Gaussian distribution centered at $I_{\text{med}}(x, y)$

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

where we set the standard deviation $\widehat{\sigma} = 5$.

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

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

For the comparative experiments in this paper we set $T = 5$ (i.e., 5 SD).

To extract the final ROIs, we apply a 5×5 dilation operator to the raw background-subtracted image D and employ a connected components algorithm. Any regions with a size less than approximately 40 pixels are discarded. A ROI extracted with this process is shown in Fig. 3(a). Notice the large number of pixels included in the ROI due to the halo.

3.2. Contour Detection

We next examine each ROI individually in an attempt to separate the foreground objects from the surrounding halo by extracting the object contours. From our earlier observations regarding thermal halos (Section 1), the gradient strengths within the ROI can be used to identify much of the object boundary.

We introduce a novel method for combining the input and background gradient strengths into a *Contour Saliency Map* (*CSM*) such that only the foreground gradients are preserved. The CSM is computed for each ROI 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

$$\text{CSM} = \min\left(\frac{\|\langle I_x, I_y\rangle\|}{\text{Max}}, \frac{\|\langle (I_x - BG_x), (I_y - BG_y)\rangle\|}{\text{Max}}\right) \tag{5}$$

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 object boundary.

The motivations for the formulation of the CSM are that (1) large non-object gradient magnitudes in the input image are suppressed (as they have small input-background gradient-difference magnitudes), and (2) large non-object input-background gradient-difference

magnitudes resulting from the halo are suppressed (as they have low gradient magnitudes in the input image). Thus, the CSM only preserves the gradients in the input image that are both strong *and* significantly different from the background. To illustrate these points, we show plots of the gradient magnitude profiles for a slice of the input and background regions in Fig. 2(d), and show the corresponding input-background gradient-difference magnitude profiles in Fig. 2(e). In Fig. 2(f), we show the resulting CSM profiles. Notice that the CSM preserves only those gradients that have a high response in *both* Fig. 2(d) and (e). The points corresponding to the left and right boundaries of the people are more clearly differentiable in the CSM than they are in either of intensity, gradient, or gradient-difference profiles. We show the creation of a CSM for an entire ROI in Fig. 3. The gradients were calculated using 7×7 Gaussian derivative masks.

In previous work (Davis and Sharma, 2004a), we suggested to create the CSM by *multiplying* the normalized input gradient magnitudes with the normalized input-background gradient-difference magnitudes. However we found that the multiplication operation tends to suppress important details and requires an additional CSM amplification step to alleviate the suppression of useful gradient information. The use of the min operator was determined to give better results and does not require the amplification step.

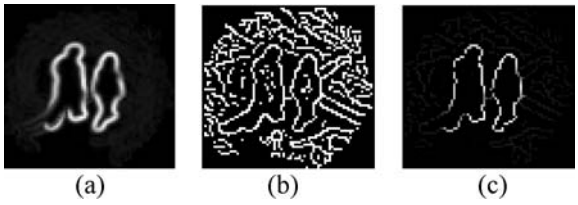

Figure 4. CSM thinning. (a) CSM. (b) Non-maximum suppression of input gradient magnitudes. (c) tCSM.

Our next step is to produce a thinned representation of the CSM, which we call the tCSM. As the CSM does not represent a true gradient image (but in fact is a *combination* of different types of gradients), 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 results of the input gradients as a thinning mask for the CSM. In Fig. 4 we show a CSM, the non-maximum suppression thinning mask (derived from the input gradients in Fig. 3(b)), and the final tCSM computed from the multiplication of the CSM with the thinning mask.

Having thinned the CSM into the tCSM, we now need to threshold the tCSM into a binary image to select the most salient contours. Our goal is to choose a single threshold that selects the majority of the object contour(s) while removing the background noise fragments. To motivate our thresholding approach, we make the following observations. Object regions, especially person regions, can be at varying temperature differentials with the environment. When the thermal variation *within* an object is much smaller than the difference in temperature *between* the object and the environment, the object pixels in the ROI are typically over-saturated and distinctly unimodal (see Figs. 3(a) and 6(a)). In other words, the object appears uniformly brighter/darker in the image than the surrounding halo (of opposite polarity). When the object-environment temperature difference is not much larger than the thermal variation within the object, the halo effect is typically weak and the object pixels in the ROI tend to be multimodal (see Fig. 5(a)). Different camera gain settings can also affect the level of saturation or modalness.

In Davis and Sharma (2004a) we proposed a simple thresholding approach where each tCSM was clustered

Figure 5. Multimodal contour selection. (a) ROI. (b) tCSM. (c) B_2. (d) B_3 (selected).

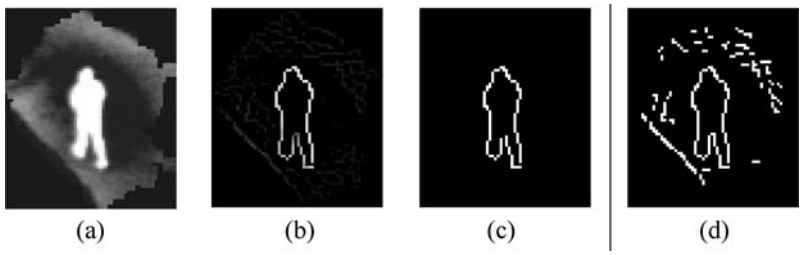

Figure 6. Unimodal contour selection. (a) ROI. (b) tCSM. (c) B_2 (selected). (d) B_3.

(using K-means) into 3 groups (low, medium, and high saliency) and thresholded by setting pixels in the lowest cluster to 0 and the remaining pixels to 1. However, as our observation regarding the typical nature of ROIs suggests, the medium cluster may or may not contain enough object contour pixels to justify its inclusion in the binary result. Specifically, when the object regions are unimodal, clustering the tCSM into 2 (rather than 3) groups and discarding the pixels in the lower cluster could produce a better result. Thus if we could reliably determine *a priori* whether the object pixels in a ROI are unimodal or multimodal, we could then choose the appropriate number of clusters (2 or 3).

Our approach is to generate two thresholded tCSMs, one result (B_2) using 2 clusters and the other result (B_3) using 3 clusters, and then to evaluate which binary image is optimal (best satisfying our goal of detecting most of the object contours without including any background noise). To rank the two binary images, we formulate a novel quality measurement Q using *average contour length* (*ACL*) and *coverage* (C). An optimally thresholded tCSM should contain mostly long contours (of the object). If the threshold is *lower* than optimal, the presence of background noise, in the form of small contour fragments, would lower the *ACL*. Similarly, if the threshold were *higher* than optimal, the existing long contours would be broken into smaller fragments and the *ACL* would be smaller than that of the optimally thresholded tCSM. An optimally thresholded tCSM would also sufficiently "cover" the ROI. The average distance of the ROI perimeter pixels to the closest pixel in the thresholded tCSM gives a quantitative measure of the coverage C of the binary result. The larger the average distance (from the perimeter pixels), the *poorer* the coverage.

Given a tCSM, we generate the two contending binary images B_2 and B_3 (obtained after clustering the non-zero pixels in the tCSM into 2 and 3 clusters respectively, and setting the pixels in the lowest cluster to 0 and the remaining pixels to 1). We then measure the quality Q of B_2 and B_3 using

$$Q(B_i) = (1-\alpha)\cdot\left(\frac{ACL(B_i)}{\max(ACL(B_2),\, ACL(B_3))}\right) + \alpha\cdot\left(1-\frac{C(B_i)}{\max(C(B_2),\, C(B_3))}\right) \tag{6}$$

The binary image that maximizes Q is chosen as the best thresholded result. Essentially, Q is a weighted sum of the normalized *ACL* and coverage values. The weighting factor α determines the 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 *ACL*s 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 *ACL*s

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

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

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

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

In Fig. 5(a) we show a ROI with multimodal person pixels (of three people), and in Fig. 5(b) we show the corresponding tCSM. The competing binary thresholded images B_2 and B_3 are shown in Fig. 5(c) and Fig. 5(d), respectively. The resulting quality values (using Eq. 6) for the images are $Q(B_2) = 0.103$ and $Q(B_3) = 0.255$. Hence, B_3 (Fig. 5(d)) with the higher quality value was correctly selected as the better thresholded result for the tCSM (as expected due to the multimodal nature of the people). The dominant factor in determining the quality for this example was the coverage

since the two *ACL*s were almost identical. We show a different ROI with a unimodal person in Fig. 6(a) and the corresponding tCSM in Fig. 6(b). The competing binary images (B_2 and B_3) are shown in Figs. 6(c) and 6(d). 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 (Fig. 6(c)) was selected as the correct thresholded image. In this example, the *ACL* was the dominating factor in the quality evaluation.

3.3. Contour Completion and Closing

If the selected binary image (B_2 or B_3) for the thresholded tCSM is guaranteed to have unbroken contours around the object (with no gaps or fragments), then a simple flood-fill operation could be used to generate the desired silhouettes. However, the contours are often 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 we can apply the flood-fill operation. Our approach is to first complete any gaps using a new search algorithm to grow out from each gap endpoint and find another contour pixel. Next, we ensure that all contours in the figure are closed. Lastly, the result is flood-filled to produce the silhouettes. To limit the search space and constrain the solution to have meaningful path completions/closings, we make use of the watershed transform.

3.3.1. Watershed Segmentation. The watershed transform (WT) is a powerful mathematical morphology tool for segmenting images by partitioning image regions with watershed lines (Couprie and Bertrand, 1997; Vincent and Soille, 1991). When computing a WT, the image is considered as a topological relief where the elevation is proportional to the pixel values. The determination of the watershed lines from this elevation surface can be described in terms of both topology (Couprie and Bertrand, 1997) and immersion simulations (Vincent and Soille, 1991). In terms of topology, a watershed line is intuitively described as "a set of points where a drop of water, falling there, may flow down towards several catchment basins of the relief" (Couprie and Bertrand, 1997). In relation to immersion, a hole is pierced at every regional minima (one in each basin) and then watershed lines are built as dams to prevent water from mixing between basins as the surface is lowered into water. We employ the efficient Vincent and Soille immersion approach (1991) as implemented in Matlab.

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. Recall that the thinned binary mask used on the CSM to create the tCSM was computed from the input gradient image. Hence we can use the WT of the same input gradient image to provide a meaningful completion guide to connect any broken contours in the binary thresholded tCSM. As long as there exists *some* information of the object boundaries in the input gradient image, the WT will produce lines along these boundaries. An over-segmented result is typical of the WT, and is normally a source of concern for most algorithms. However, our application of the WT is not segmentation. We instead employ the WT as a novel method to limit the search space and provide a meaningful guide for contour completion. In such a scenario, an over-segmented result still yields a much smaller number of paths than *all* possible paths, hence providing a limited set of good path candidates to choose from when closing a gap. Also, due to the small size of the ROIs (as compared to the much larger size of the entire image), the cost of computing the WT for a ROI is fairly minimal. In Fig. 7, we show a ROI,

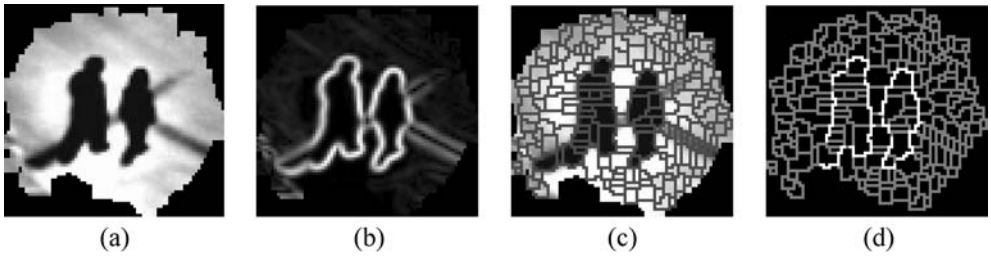

Figure 7. Watershed analysis. (a) ROI. (b) Input gradient magnitude. (c) WT overlaid on ROI. (d) Thresholded tCSM overlaid on WT.

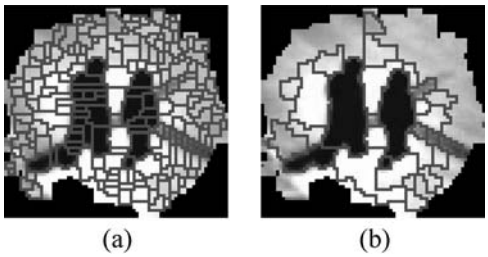

Figure 8. Basin merging. (a) Original WT overlaid on ROI. (b) Merged WT overlaid on ROI.

the corresponding input gradient magnitude, the WT of the gradient magnitude (overlaid on the ROI), and the thresholded tCSM overlaid on the WT lines.

Apart from providing limited, yet the most likely, completion paths for the tCSM, the WT also enables us to eliminate small, stray contour fragments (barbs) that may harm the completion/closing routines. To remove these barbs, we first merge the watershed basins into a coarser segmentation of the ROI, and then examine the length of the contour segments in relation to the basin boundaries.

Our 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 gives us a lower resolution segmentation of the ROI. We shown an example before and after basin merging in Fig. 8. Other merge algorithms could also be applied (Najman and Schmitt, 1996; Lemaréchal and Fjørtoft, 1998).

We use this coarser resolution WT to validate the contour segments of the thresholded tCSM to eliminate any small noisy barbs that might exist. Based on the merged WT, the thresholded tCSM is partitioned into distinct segments that divide pairs of adjacent basins. A tCSM 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. The intuition behind the process is that at least half of the boundary between two neighboring regions must be reinforced, otherwise the tCSM pixels on the boundary are likely to be noise. We show a threhsolded tCSM, the image overlaid on the merged WT, and the result after the validation process in Fig. 9. Notice that several small fragments are removed after the validation process. The merged WT is used for only this step (to validate the contours in the tCSM), and the remaining completion/closing processes employ the original WT.

3.3.2. Completion and Closing. Our next step is to attempt to complete any contour gaps using the original watershed lines as plausible connection pathways for the (validated) thresholded tCSM. Each "loose" endpoint of the contour segments (found using 3×3 neighborhood analysis) is forced to grow outward (not permitted to move along its own contour segment) along the watershed lines until another contour point is reached. To find the optimal path, we employ the A* search algorithm (Russell and Norvig, 2003) that minimizes the expected cost *through* the current pixel location to reach another contour point. The Euclidean distance from the current pixel location to the location of remaining thresholded tCSM contour pixels is employed 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.

The contour completion process is very effective in correctly closing a large proportion of the gaps between

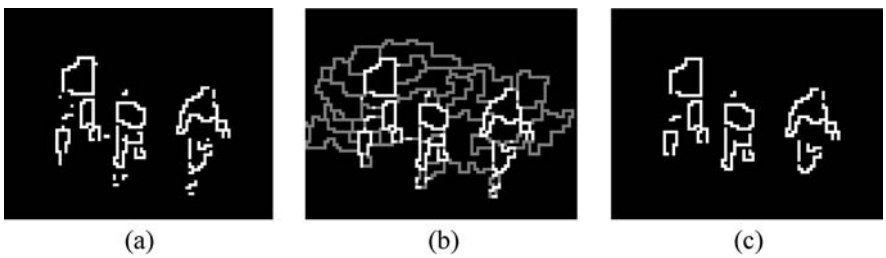

Figure 9. Contour validation. (a) Thresholded tCSM. (b) Thresholded tCSM overlaid on merged watershed lines. (c) Thresholded tCSM after contour validation.

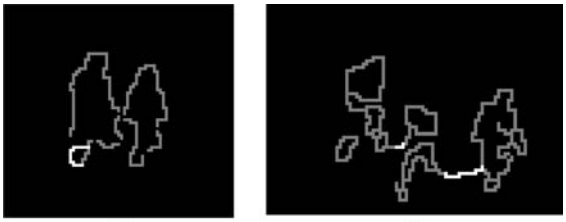

Figure 10. Contour completion problem scenarios (before analysis and correction). (a) Incorrect loop-back. (b) False connection between different people.

contour segments. However, due to it's simplicity, not every path that is created produces a reasonable completion. Two problem scenarios can occur: (1) a path grows outward but loops around back to the pixel's own contour segment (see Fig. 10(a)), and (2) a path forms between (and connects) two different objects/people (see Fig. 10(b)).

We first address the initial problem scenario. Let P_1 be the path created by joining a contour endpoint to some other contour pixel. We attempt a new path P_2 between the same pair of points, but this time we allow the path to grow back along its own contour to reach the destination pixel. If the two paths are equivalent, we keep P_1. If the paths are different, we choose the path having the maximum support from the tCSM. We first compute the amount of support for a path by counting the number of pixels n in the *unthresholded* tCSM (i.e., the thinned CSM) that exist along that path. We use the unthresholded tCSM to account for any pixels that may have been incorrectly deleted during the thresholding and validation process. We then choose between P_1 and P_2 using

$$P = \begin{cases} P_1 & \dfrac{|n_1|}{|P_1|} > \dfrac{|n_2|}{|P_2|} \\ P_2 & \text{otherise} \end{cases} \tag{9}$$

This method removes the false loop-back in Fig. 10(a).

The second problem scenario typically happens when a short contour fragment juts outwards from the object boundary (which may still occur due to small basin sizes in spite of the merging process in the prior validation step). Only those new paths that do not form a closed loop (as in Fig. 10(b)) are examined in this step (closed loops are acceptable completions). For a given path, we first assign to each pixel along the path the maximum saliency value (within its 3×3 neighborhood) from the original CSM. We note that since each path must start and terminate at a selected contour pixel, the two ends of a path are at a local maxima in the saliency values. Then the saliency profile of the entire path is examined. If a distinct valley is observed in the saliency values, the path is declared invalid. The presence of a valley indicates that the path was likely constructed over pixels of low saliency (non-object contour pixels), and thus the path should be deleted. The following heuristic is employed to quickly determine the presence or lack of a valley. Given a list of saliency values S for a path, the locations (M_l, M_r) of the leftmost and rightmost maxima are determined by seeking inward on both sides from the path endpoints (the maxima are typically the endpoints). A valley is deemed *not* present only if there exists a local maxima between M_l and M_r with a value greater than $S(M_l)$ or $S(M_r)$, or if $M_l = M_r$ (for the case of S linearly increasing or decreasing between the endpoints). The two incorrect paths in Fig. 10(b) are removed with this approach.

The contour completion and verification processes described above are repeated in tandem until no new paths are found. Lastly, when all gaps are completed, we perform a final match-consistency check. If endpoint E_1 has grown to some non-endpoint, and another endpoint E_2 has grown to E_1, we favor the $E_1 - E_2$ connection and remove the path from E_1 to the non-endpoint. We show a (validated) thresholded tCSM in Fig. 11(a) and the completion result in Fig. 11(b). The

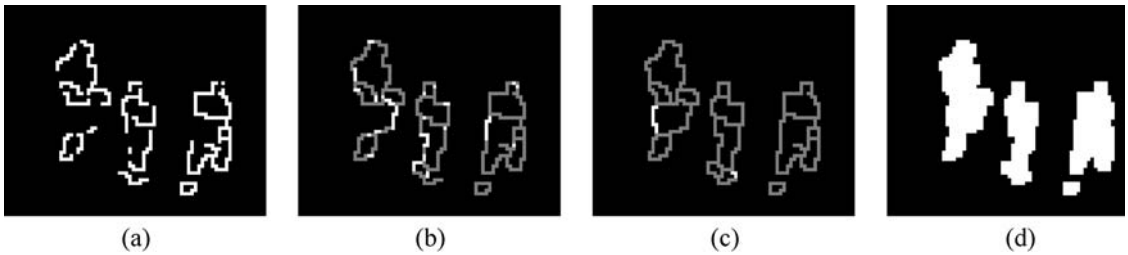

Figure 11. Contour completion, closing, and flood-filling pipeline. (a) Original thresholded tCSM. (b) Completed contour result (white lines are new paths). (c) Closed result of (b) (white lines are new paths). (d) Flood-filled silhouettes.

uncompleted contour in the foot region of the middle person was due to the verification step where an alternate path had more tCSM support. This contour segment will be addressed in the following closing operation.

In the final stage, we ensure that every contour in the image is part of a closed loop (required for flood-filling). We consider any "loose" contour end fully contained in the interior of a closed loop as part of the loop, and thus that endpoint does not need to be closed (that contour fragment could be removed). In geometric terms, our goal is a binary contour image that is equivalent to the closure of its interior. First, all those contours not part of a closed loop are identified (by region growing along the contours). Either a contour is part of a closed loop or it connects other closed loops. A contour could additionally have an external "loose" end (not connected to another loop/contour), as seen in the foot region of the middle person in Fig. 11(b).

Given an un-closed contour C, the set of closed contour loops (L) connected by C are identified. Two loops, L_1 and L_2, are chosen at random from L, and we select the points p_1 and p_2 at which the Euclidean distance between L_1 and L_2 is the least (p_1 and p_2 may not be unique, and all pairs could be examined if desired). To close the region, we require a path connecting p_1 and p_2 *other than the contour* C. To find the solution that creates the minimum number of new contour pixels on the watershed lines, we give no penalty (step cost) in the A* algorithm for moving along existing contour pixels on the watershed (allowing a "free glide" along existing contour pixels).

In Fig. 12(a) we present an example of a typical scenario, where the contour C is shown in gray, the loops L_1 and L_2 connected by C are shown in white, and the points, p_1 and p_2, are circled. Our approach is to trace two paths, one from p_1 to p_2 and the other from p_2 to p_1, using the A* search strategy along the watershed lines. The two paths may be different due to the "free glide" allowed along existing contours. To ensure that the new paths are distinct from C, we ignore and block off the pixels of C that are closest to L_1 and L_2 during the A* search (marked dark gray in Fig. 12(a)). We choose the path that maximizes the interior area of the figure. In the unlikely situation when no possible path exists between p_1 and p_2 (no watershed path), we default to a direct straight-line closing.

In the alternate case when C has loose/open ends, one of them is chosen at random as the starting point for the A* search, and all pixels belonging to the loops in L are considered valid destination points. In this case, we ignore all pixels on C within a 3×3 neighborhood of the starting point to ensure that the new path is distinct from C.

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 Figs. 11(c) and 12(b). After the completion and closing procedures, a simple flood-fill operation is employed to create the silhouettes. We present the final flood-filled silhouettes for the closing result of Figs. 11(c) in 11(d).

4. Experiments

To examine our contour-based background-subtraction approach, we tested the method with six challenging thermal video sequences recorded at very different environmental conditions. Quantitative results are provided comparing our method with three other approaches using a subset of images with manually labeled person regions. We also show the feasibility of our results for tracking using a blob correspondence and tracking approach over multiple frames in a difficult video sequence. Lastly, we examine the approach for other non-person object classes and demonstrate the applicability of our approach to color imagery.

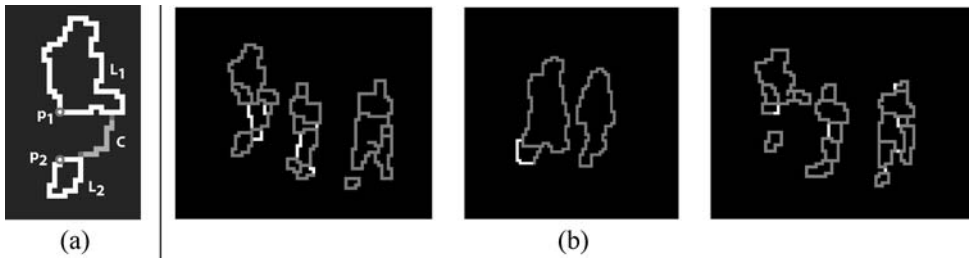

Figure 12. Contour closing. (a) Open contour connecting two loops. (b) Examples of closing mechanism (paths shown in white).

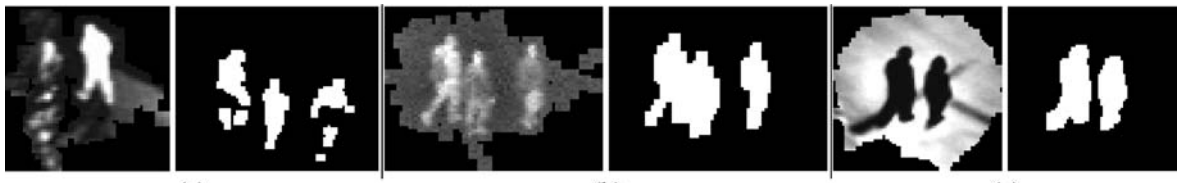

Figure 13. Difficult ROIs. (a) Deletion of body regions. (b) Joining multiple people. (c) Silhouette extension into the background.

The thermal dataset was recorded on the campus of Ohio State University in different seasons (Winter, Summer) showing pedestrians at various times-of-day (morning, afternoon, evening). The number of frames in each sequence was Sequence-1:297, Sequence-2:314, Sequence-3:1466, Sequence-4:500, Sequence-5:1000, and Sequence-6:1001. Three sequences (1–3) were recorded from low vantage points (3- and 4-story elevations) and three sequences (4–6) were recorded from the roof of an 8-story building. The sequences were captured using two different Raytheon ferroelectric BST sensor cores (300D, 250D) with a sensor array size of 320×240. The sizes of the input sequences processed were half-resolution at either 320×240 or 360×240 (depending on the NTSC or DV recording source). Example images from this dataset are shown in Fig. 14.

To demonstrate the generality and applicability of our approach, we extracted silhouettes from each video sequence with our proposed method using the **same parameter/threshold settings for all sequences** (no individual parameter tuning for each sequence). To give flexibility to a human operator (e.g., for human-in-the-loop surveillance monitoring) to select/show only the most confident detections, we weighted each resulting silhouette in the image with a contrast value $\mathcal{C}$ calculated from the ratio of the maximum input-background intensity difference within the silhouette region to the full intensity range of the background image BG

$$\mathcal{C}(sil) = \frac{|\max(I(sil)) - \max(BG(sil))|}{\max(BG) - \min(BG)} \quad (10)$$

where sil represents a particular silhouette region detected in the input image I. A final user-selected threshold on $\mathcal{C}$ could easily be used to remove any minimal-contrast (noise) regions.

Our results showing the contrast-weighted silhouettes (using Eq. 10) for the images in Fig. 14 are shown in Fig. 15. The results demonstrate the overall ability of our algorithm to extract silhouettes in very different imagery. Note that even though small ROIs (<40 pixels) are removed, small valid-sized ROIs that *do not* contain a foreground object (person) may result in small silhouette regions (of noise). However, these regions typically have a low contrast value and could be easily removed using a threshold on $\mathcal{C}$.

For Sequence-1, the algorithm was able to detect considerable portions of the people despite the very low person-background thermal differences and low gradients. Additionally, a small animal in the top-right corner was detected in several images. In spite of the thermal similarity of the cross-walk and people in Sequence-2, the silhouettes were extracted and separated quite well. The small fragmented silhouettes detected in the top-left corner of Sequence-3 were due to the people being partially occluded by tree branches. The people in Sequence-4 through Sequence-6 were far from the camera and at various thermal contrasts with the environment, but their silhouettes were detected reasonably well. In several of the images for different sequences, the approach was able to successfully distinguish multiple people in close proximity (within the same ROI).

There are three general problems that can occur that deserve mentioning. First, strong thermal similarity between regions of the foreground object and the background can result in portions of the object being deleted, as shown in Fig. 13(a). Second, if objects in the ROI are in very close proximity, foreground-background similarity can result in incorrect joining of the silhouettes. This case is shown in Fig. 13(b). Lastly, when certain boundaries of the foreground objects have low saliency, the silhouette can be extended into the background region. As shown in Fig. 13(c), the thermal intensity of the people is similar to the background cross-walk on the pavement and hence the low contour saliency at the overlapping pixels resulted in a silhouette with the leftmost foot region extending slightly into the background. Since all three difficulties arise due to

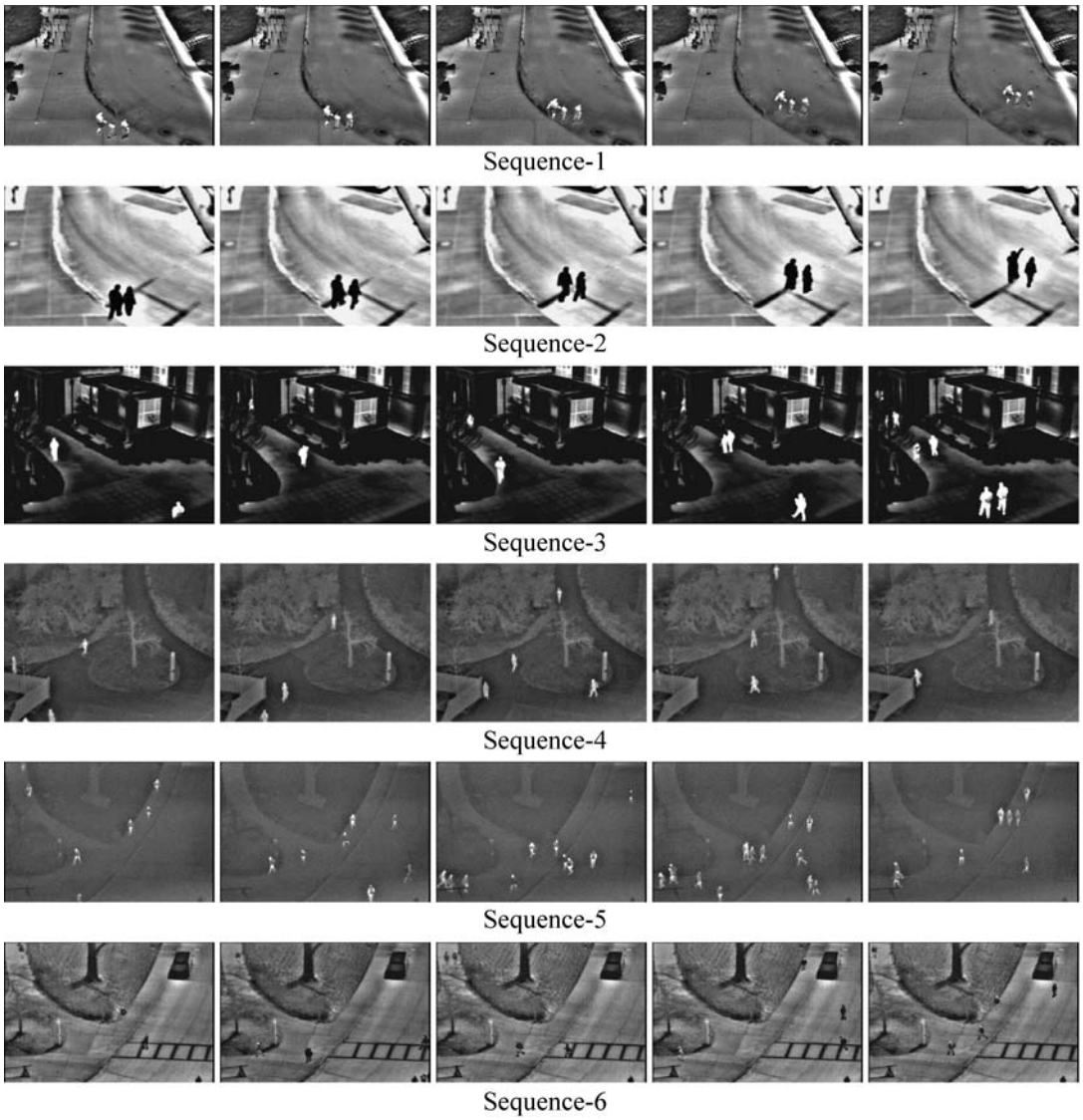

Figure 14. Example images from the thermal imagery dataset.

insufficient contrast between the foreground and background in the ROI, most other intensity-based (thermal/grayscale) methods would have similar problems.

4.1. Quantitative Comparison

In order to quantitatively measure the performance of the detection results (and to compare with other approaches), we collected a subset of 30 images from the dataset (5 images spanning each of the 6 sequences) and manually segmented the person regions. The selected images are shown in Fig. 14. Five people familiar with thermal imagery were independently asked to hand-draw silhouettes on the people (marked *a priori* with a bounding boxes) appearing in each of the 30 images. The instructions were as follows:

> *"In each of the following images, boxes have been drawn around regions containing people. Using the 'pencil' tool in Adobe Photoshop, mark all regions belonging to the people within each box. Consider all clothing accessories (e.g., hats) and other carried objects (e.g., backpacks) as belonging to the person region. Regions belonging to a single person should not be disconnected. You may zoom in/out of the image as needed."*

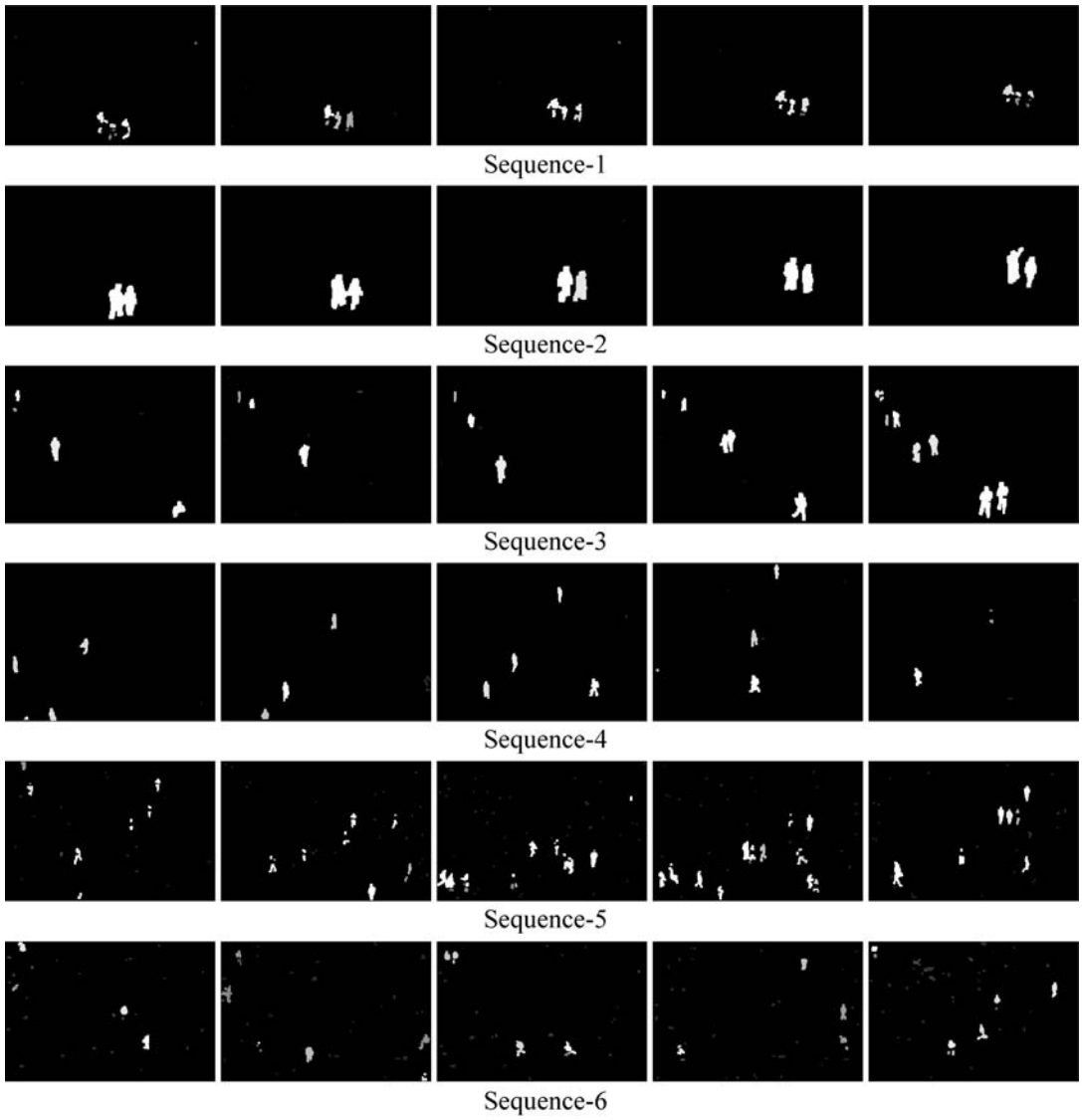

Figure 15. Detection results (contrast-weighted silhouettes) for example images shown in Fig. 14.

For each of the 30 images, the median image of the 5 manually drawn silhouette images was computed. The 30 median silhouette images were used for algorithm evaluation and comparison. Six of the median silhouette images (one from each sequence) are shown in Fig. 16.

Using the manually segmented images, we compared the results of our algorithm with three alternate methods: statistical background-subtraction, image-differencing, and hot-spot detection. These simple extraction/detection approaches are commonly used in both color and thermal imagery (e.g., Wren et al., 1997; Iwasawa et al., 1997; Bhanu and Han, 2002). Statistical background-subtraction (BS) involves thresholding the Mahalanobis distance of an input image to the mean/variance background model at each pixel. We in fact use this method as our initial stage to detect the ROIs. Image differencing (ID) thresholds the absolute difference of an input image with a background image (usually a mean or median image). Hot-spot detection (HD), used specifically for thermal imagery, directly compares an input thermal image with a threshold, with the expectation that the object/person is hotter than the environment. The background models

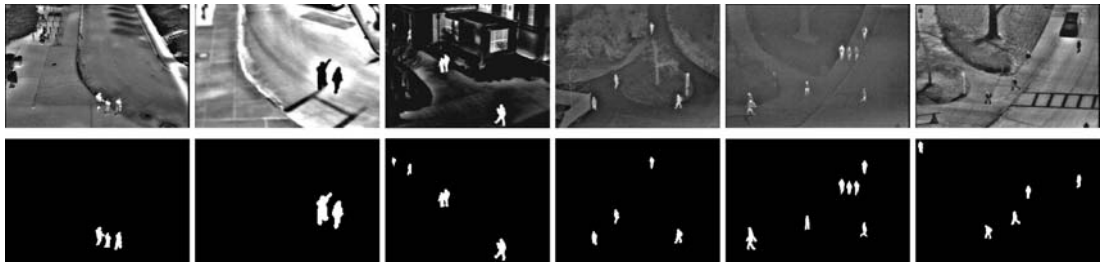

Figure 16. Examples of manually segmented silhouettes for images taken from the original dataset of 6 sequences.

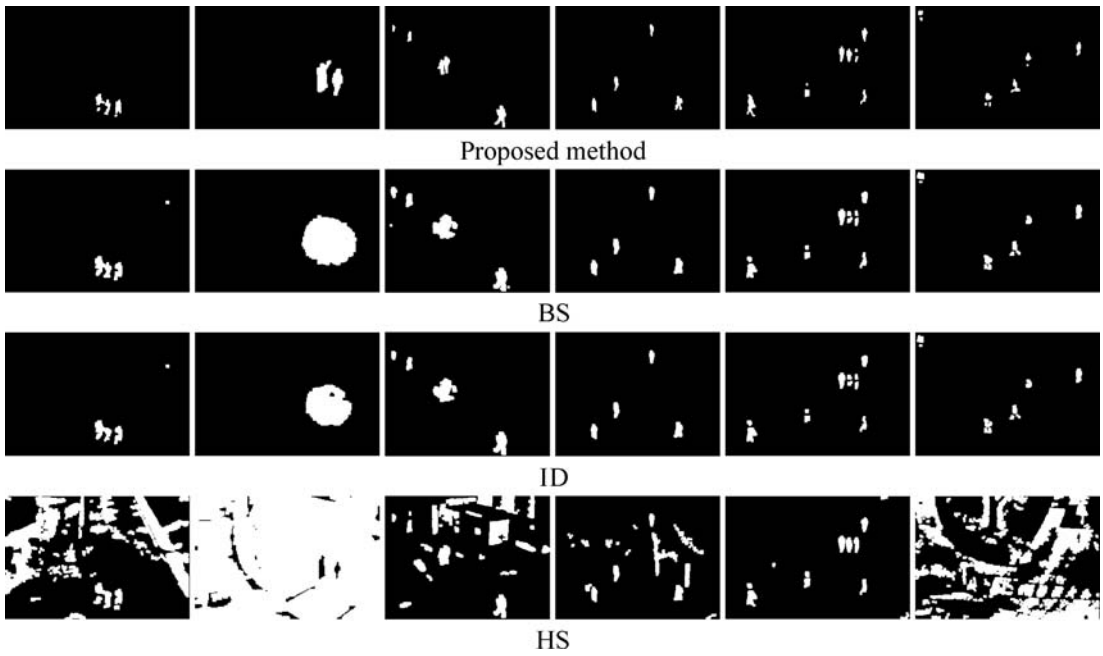

Figure 17. Visual comparison of detection results of the four algorithms across different images from Fig. 16.

for BS, ID, and our approach were computed with the technique described in Section 3.1 using the entire duration for each sequence. After thresholding in each of the alternate methods, a 5×5 dilation filter was employed, and regions less than approximately 40 pixels were removed (as used in our approach to detect the ROIs). No erosion filter was applied in any of the methods.

4.1.1. Comparison 1. To quantitatively compare the results of our algorithm with the three alternate approaches, we examined *Sensitivity* and *Positive Predictive Value* (PPV) measurements using the manually segmented images as ground truth. Sensitivity refers to the fraction of person pixels that are correctly detected by the algorithm, while PPV represents the fraction of detections that are in fact person pixels. To choose the best threshold for each of the methods, the relevant algorithm threshold was adjusted over a large range and the threshold yielding the largest sum of the Sensitivity and PPV over all of the 30 images was selected (we want both Sensitivity and PPV to be large). For our algorithm, only the final contrast threshold ($\mathcal{C}$) was adjusted (the remaining parameters were fixed for all sequences).

We show the silhouette results obtained by each of the algorithms on representative images (from the set of 30) in Fig. 17. The Sensitivity and PPV results for the 30 images are shown in Table 1. A favorable algorithm should attain high values (close to 1) for *both* Sensitivity and PPV (i.e., the best method should achieve a high detection rate with low false positives). Due to

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

Method		Sequence-1	Sequence-2	Sequence-3	Sequence-4	Sequence-5	Sequence-6	Average
Proposed	S	0.662	0.958	0.943	0.822	0.708	0.677	0.793
	PPV	0.973	0.928	0.952	0.960	0.948	0.855	0.936
BS	S	0.905	1.000	1.000	0.949	0.838	0.803	0.916
	PPV	0.702	0.278	0.463	0.614	0.698	0.644	0.567
ID	S	0.861	0.998	0.999	0.938	0.834	0.776	0.901
	PPV	0.733	0.377	0.544	0.634	0.697	0.648	0.606
HS	S	0.889	–	0.994	0.990	0.921	–	0.949
	PPV	0.034	–	0.107	0.127	0.585	–	0.213

Table 2. Average distance of non-person pixels detected (false positives) to manual silhouette shapes.

Method		Sequence-1	Sequence-2	Sequence-3	Sequence-4	Sequence-5	Sequence-6	Average
Proposed	Mean	1.026	1.064	1.012	1.088	1.063	1.710	1.1605
	SD	0.130	0.235	0.070	0.326	0.283	1.242	
BS	Mean	15.919	6.269	4.346	1.403	1.326	1.522	5.1308
	SD	44.576	3.895	9.451	0.505	0.462	0.799	
ID	Mean	15.818	4.434	2.771	1.349	1.325	1.511	4.5347
	SD	44.269	2.893	2.554	0.466	0.463	0.799	

the presence of halos in sequences 2–4, the Sensitivity of BS and ID were nearly 1 (the silhouettes covered much more than the person region), but the corresponding PPV rates were poor (<.64). The HS approach was only applicable to the images from sequences 1, 3, 4, and 5 where the person generally was brighter than the background. This fact clearly renounces HS as a persistent method. The valid HS results had Sensitivity values near 1, but had extremely low PPV rates (most near .1). Our approach performed best when the halos are the most prominent (sequences 2–4). For these sequences, our algorithm had very high Sensitivity *and* PPV. In sequences 1, 5, and 6, our approach did not extract as much of the person as did the other algorithms (as shown by the lower Sensitivity values). However we were able to detect nearly 70% of the desired silhouette mass while maintaining a higher PPV than the competing algorithms. Overall, the average results in the final column of the table demonstrate that our approach provides the best combination of the Sensitivity and PPV rates.

4.1.2. Comparison 2. Our next experiment evaluated how closely the detection results of each algorithm matched the manually segmented silhouettes. For a perfect match, the detection results should be completely contained within the manually segmented silhouettes, and every point on the silhouettes should have been detected. We computed the mean and standard deviation (SD) of the closest distances of the non-person pixels detected (false positives) to the corresponding manually segmented silhouettes for each method. Table 2 shows the values obtained from our algorithm, BS, and ID using the 30 test images. The HS approach was not evaluated as it performed too poorly in the previous experiment to be considered a viable algorithm.

In nearly all of the sequences, our method had an average distance of about 1 pixel. The other two methods were in the range of 1–16 pixels. The unusually high mean distances for BS and ID in Sequence-1 were due to a small animal detected in the top-right corner in the scene. Though this animal was initially detected by our algorithm, it was assigned a very small contrast value and was automatically thresholded out of the final results.

4.1.3. Comparison 3. We also need to measure how well the detected silhouette pixels "cover" the manual silhouettes. We computed the mean and standard deviation of the closest distances of the missed pixels (false negatives) to the correctly detected pixels (true positives). Table 3 shows these values for the three

Table 3. Average distance of undetected person pixels (false negatives) to detected person pixels (true positives).

Method		Sequence-1	Sequence-2	Sequence-3	Sequence-4	Sequence-5	Sequence-6	Average
Proposed	Mean	2.695	1.040	1.308	4.443	2.213	9.957	3.6093
	SD	2.747	0.185	0.639	5.038	1.551	19.851	
BS	Mean	1.873	1.000	1.000	6.968	1.875	14.360	4.5127
	SD	1.198	0.000	0.000	4.991	1.090	24.280	
ID	Mean	2.334	1.000	1.143	6.664	1.974	13.931	4.5077
	SD	1.869	0.000	0.378	5.233	1.184	22.798	

algorithms being compared. A high mean and standard deviation implies that the detected person regions are widely fragmented.

In Table 3, the mean distance values for sequences 1 and 5 are higher for our approach than for the other algorithms. This is consistent with the lower Sensitivity values obtained by our algorithm for those sequences. However, the mean values of BS and ID are only marginally less than ours. The results of our algorithm could be further improved by using shape information during the contour completion/closing stage. Sequences 2 and 3 have very low mean values for all the algorithms, showing that the detection results in all of the approaches adequately fill out the manually segmented silhouettes. Furthermore our algorithm performs much better than the competing algorithms in sequences 4 and 6. The unusually high mean values for Sequence-6 are because large parts of a person region in one image did not pass the corresponding threshold in any of the approaches.

To truly evaluate and compare the detection results for the different algorithms, we must examine the two distance tables together (Tables 2 and 3). The last column in the tables show that the silhouettes extracted by our algorithm match the manual silhouettes better on average than those generated by the alternate approaches.

4.2. *Extensions*

We examined the effect of different contrast thresholds ($\mathcal{C}$) on the quality of our results. The trade-off between PPV and Sensitivity for different thresholds on $\mathcal{C}$ (equally spaced between 0 and 0.55) are shown in Fig. 18, where PPV is plotted on the y-axis, and (1-Sensitivity) is plotted on the x-axis. As expected, as the threshold on $\mathcal{C}$ is raised, the PPV increases and the Sensitivity decreases. The range of thresholds that yield the best results can easily be determined from the figure.

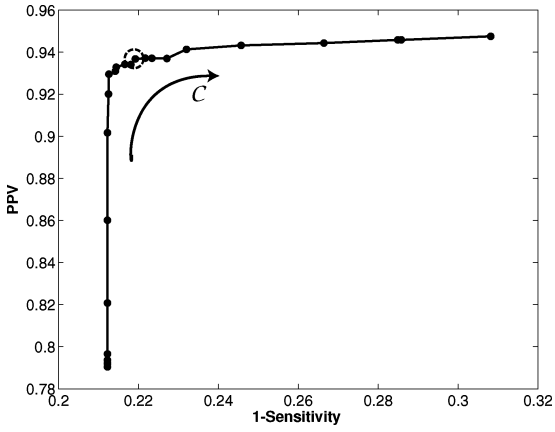

Figure 18. Trade-off between Sensitivity and PPV for different contrast thresholds $\mathcal{C}$. The threshold used in the previous experiments is circled.

The contrast threshold used in the above experiments is marked in the figure.

One of the main drawbacks of existing background-subtraction methodologies (BS, ID) is that detection results depend heavily on the chosen threshold values. While our approach also utilizes background-subtraction as an initial stage of processing, the final detection results are quite robust to different threshold

Table 4. Quality of detection results with different background-subtraction thresholds T (Eq. (4)).

T	Sensitivity	PPV
5	0.793	0.936
6	0.776	0.944
7	0.772	0.941
8	0.752	0.946
9	0.742	0.949
10	0.711	0.948

Figure 19. Results of blob-level tracking for Sequence-1.

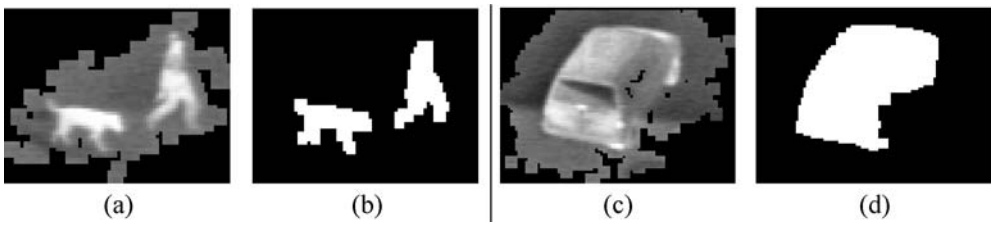

Figure 20. Extraction of non-person object classes. (a) ROI of dog and person. (b) Final silhouettes. (c) ROI of vehicle. (d) Final silhouette.

values. In Table 4 we show the best Sensitivity and PPV measures (determined from multiple contrast C values) obtained by our algorithm on the set of 30 images for different values of the initial background-subtraction threshold T in Eq. (4). As the Sensitivity and PPV numbers indicate, our algorithm performs consistently well over a fairly wide range of background-subtraction thresholds.

To further demonstrate the usability of our detection results, we tested a blob-level tracking method with one of our more fragmented silhouette sequences (Sequence-1). The blob tracker is based on the work of Masoud and Papanikolopoulos (2001), which uses the overlap-area of blobs in successive frames to create an association graph for consistently tracking people even when blobs split and/or merge. The tracking was manually initialized by identifying the blobs belonging to each person in the first image of the sequence. The three people in Sequence-1 were correctly identified and tracked throughout the entire sequence. Images from the tracking result are shown in Fig. 19 (we did not track the infrequent appearance of the animal in the upper-right corner). As we can see from the images, even though the detected blobs are fragmented, they can still be consistently grouped and tracked over time.

Although we evaluated the performance of our algorithm for foreground regions of people, the algorithm can be used to extract other object classes as well. In Fig. 20 we show the silhouettes extracted for two very different object classes. Figure 20(a) shows a ROI containing a person leading a dog on a leash, and Fig. 20(b) shows the corresponding silhouettes.

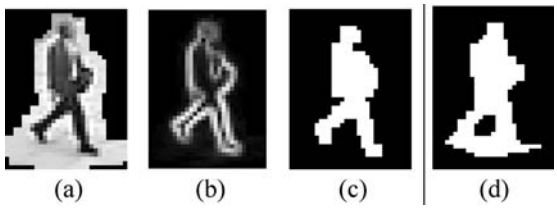

Figure 21. Detection results in color imagery. (a) Input ROI. (b) CSM. (c) Final silhouette. (d) Background-subtraction result.

As can be seen, the completion strategy is able to generate well-separated and reasonably shaped silhouettes of the person and the dog. In Fig. 20(c) we show a ROI containing a vehicle and in Fig. 20(d) we present the extracted silhouette. We note that the completion strategy could conceivably employ heuristics to bias the method toward specific types of foreground objects (if desired). For example, to better segment vehicles from the background (e.g., in Fig. 20(c)), the contour completion phase could be limited to the use of longer straight lines and T-junctions for the connections.

Though our approach was motivated by the presence of halos in thermal imagery produced by ferroelectric BST sensors, the method is also applicable to the color domain. In Fig. 21 we show results obtained from images recorded using a standard CCD color camera. The color-space was first converted from *RGB* to *YCbCr*, and only the intensity component (Y) was processed. Figure 21(a) shows an example ROI of a person. Note that the ROI includes a soft diffused shadow cast by the

person. In Fig. 21(b) we show the corresponding CSM and Fig. 21(c) shows the final silhouette extracted with our approach. We used the same parameters/thresholds in our algorithm for color imagery as we did for the experiments on thermal imagery. We see that our approach is quite applicable to the visible domain. As a comparison, we also show the results obtained from standard background-subtraction (BS) in Fig. 21(d). The background-subtraction threshold was fine-tuned over a number of images to ensure that all person regions were detected while minimizing the detection of non-person regions (the resulting threshold was 15 SD). Comparing Fig. 21(c) and (d), we see that our gradient-based approach is capable of eliminating soft diffused shadows in color imagery, which are incorrectly detected by traditional background-subtraction (which would require an additional shadow-removal step).

The frame-rate of the proposed method depends considerably on the number of ROIs (foreground objects) in the images. The most computationally-intensive components of the algorithm, beyond the initial background-subtraction process, are the search and validation methods of the contour completion and closing procedures. Thus the number of ROIs and the number of objects within each ROI (and their modality) will directly influence the amount of time spent in this stage of the algorithm. The watershed transform (Vincent and Soille, 1991) can also be costly to compute, but as it is only applied in small ROIs (rather than the entire image), it does not significantly contribute to the overall processing time. Using un-optimized Matlab code on a 2.8 GHz Pentium 4 computer, we experienced typical processing times of .97 to 4.79 seconds per ROI depending on the complexity of the ROI (number of people, modality).

5. Summary

We presented a new contour-based background-subtraction method to extract foreground objects in thermal imagery over a wide range of environmental conditions (including day/night and Winter/Summer scenarios). Our approach was designed to handle the problems typically associated with thermal imagery produced by common ferroelectric BST sensors such as halo artifacts and uncalibrated polarity switches. These problems typically limit the applicability of standard object extraction/detection algorithms, such as statistical background-subtraction, image-differencing, and hot-spot detection, to only certain types of thermal imagery.

Our approach first uses statistical background-subtraction to identify local regions-of-interest containing the foreground object and the surrounding halo. The input and background gradient information within each region are then combined into our novel Contour Saliency Map (CSM) representation. The CSM is thinned using a non-maximum suppression mask of the input gradients. The most salient contours are then selected using a new thresholding strategy based on competitive clustering. Any broken contour fragments are completed and closed using a new watershed-constrained A* search strategy. The final contours are then flood-filled to produce silhouettes. Lastly, a contrast value is assigned to each silhouette to enable a final selection threshold.

Experiments with our method and six challenging thermal video sequences of pedestrians recorded at very different environmental conditions showed promising results. We demonstrated the generality of the approach using a single set of parameters/thresholds across the dataset. A manually segmented subset of 30 images was used to compare the results of our algorithm with statistical background-subtraction, image-differencing, and hot-spot detection. Quantitative results using Sensitivity, Positive Predictive Value, and false positive/negative distances demonstrated the enhanced performance of our approach over the other methods. The detected person pixels were found to coincide fairly well with the manually segmented silhouettes. Our approach was more robust than the other methods across different environmental conditions which created large variations in the thermal imagery. As the approach is not limited to only extracting silhouettes of people, we also demonstrated the method for extracting silhouettes of a dog and vehicle. Furthermore, we showed the applicability of the approach to color imagery.

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. Of special interest will be to employ the extracted silhouettes in an activity recognition system (e.g., Bobick and Davis, 2001) for event detection. We expect our approach to be an effective advancement towards persistent and automatic video surveillance.

Acknowledgments

This research was supported in part by the National Science Foundation under grant No. 0236653, the Secure Knowledge Management Program, Air Force Research Laboratory (Information Directorate, Wright-Patterson AFB, OH), and the U.S. Army Night Vision Laboratory.

References

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

Bhanu, B. and Holben, R. 1990. Model-based segmentation of FLIR images. *IEEE Trans. Aero. and Elect. Sys.*, 26(1):2–11.

Bobick, A. and Davis, J. 2001. The recognition of human movement using temporal templates. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 23(3):257–267.

Couprie, M. and Bertrand, G. 1997. Topological grayscale watershed transformation. In *Vision Geometry V*, Vol. 3168, SPIE, pp. 136–146.

Cutler, R. and Davis, L. 1999. Real-time periodic motion detection, analysis, and applications. In *Proc. Comp. Vis. and Pattern Rec.*, IEEE, pp. 326–332.

Danker, A. and Rosenfeld, A. 1981. Blob detection by relaxation. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 3(1):79–92.

Davis, J. and Keck, M. 2005. A two-stage template approach to person detection in thermal imagery. In *Proc. Wkshp. Applications of Comp. Vis.*

Davis, J. and Sharma, V. 2004a. Robust background-subtraction for person detection in thermal imagery. In *IEEE Int. Wkshp. on Object Tracking and Class. Beyond the Vis. Spect*.

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

Elgammal, A., Harwood, D., and Davis, L. 2000. Non-parametric model for background subtraction. In *Proc. European Conf. Comp. Vis.*, pp. 751–767.

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

Haritaoglu, I., Harwood, D., and Davis, L. 1998. W4: Who? When? Where? What? A real time system for detecting and tracking people. In *Proc. Int. Conf. Auto. Face and Gesture Recog.*, pp. 222–227.

Harville, M. 2002. A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. In *Proc. European Conf. Comp. Vis.*, pp. 543–560.

Hoist, G. 2000. *Common Sense Approach to Thermal Imaging*. SPIE Press, Bellingham, Washington.

Iwasawa, S., Ebihara, K., Ohya, J., and Morishima, S. 1997. Real-time estimation of human body posture from monocular thermal images. In *Proc. Comp. Vis. and Pattern Rec.*, IEEE, pp. 15–20.

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

Kummer, S. 2003. The eye of the law. *OE Magazine*, 3(10):22–25.

Lemaréchal, C. and Fjørtoft, R. 1998. Comments on geodesic saliency of watershed contours and hierarchical segmentation. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 20(7):762–763.

Lipton, A., Fujiyoshi, H., and Patil, R. 1998. Moving target classification and tracking from real-time video. In *Proc. Wkshp. Applications of Comp. Vis.*

Masoud, O. and Papanikolopoulos, N. 2001. A novel method for tracking and counting pedestrians in real-time using a single camera. *IEEE Trans. on Vehicular Tech.*, 50(5):1267–1278.

Mittal, A. and Paragios, N. 2004. Motion-based background subtraction using adaptive kernel density estimation. In *Proc. Comp. Vis. and Pattern Rec.*, pp. 302–309.

Monnet, A., Mittal, A., Paragios, N., and Ramesh, V. 2003. Background modeling and subtraction of dynamic scenes. In *Proc. Int. Conf. Comp. Vis.*, pp. 1305–1312.

Najman, L. and Schmitt, M. 1996. Geodesic saliency of watershed contours and hierarchical segmentation. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 18(12):1163–1173.

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

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

Pandya, N. and Anda, J. 2004. Across the spectrum. *OE Magazine*, 4(9):28–31.

Russell, S. and Norvig, P. 2003, (eds.) *Artificial Intelligence: A Modern Approach.* Prentice Hall.

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

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

Vincent, L. and Soille, P. 1991. Watershed in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 13(6):583–598.

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

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

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

Yilmaz, A., Shafique, K., and Shah, M. 2003. Target tracking in airborne forward looking infrared imagery. *Image and Vision Comp.*, 21(7):623–635.

Zhong, J. and Sclaroff, S. 2003. Segmenting foreground objects from a dynamic, textured background via a robust kalman filter. In *Proc. Int. Conf. Comp. Vis.*, pp. 44–50.