Robust Background-Subtraction for Person Detection in Thermal Imagery*

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 to detect people in widely varying thermal imagery. Statistical background-subtraction is first used to identify local regions-of-interest. Within each region, gradient information in the foreground and background are combined to form a contour saliency map. After thinning, an A path-constrained search along watershed boundaries is used to complete any broken contour segments. Lastly, the contour image is flood-filled to produce silhouettes. Results are presented that demonstrate the robustness of the approach to detect people across a wide range of thermal imagery using a fixed set of parameters.*

1. Introduction

We present a new background-subtraction technique to robustly detect people in thermal video across different environmental conditions. Thermal video cameras detect the amount of thermal radiation emitted/reflected from objects in the scene, and are applicable to both day and night scenarios. Therefore, they are a prime candidate for a persistent (24-7) video system for surveillance and monitoring. As long as the thermal properties of a person are slightly different (higher or lower) from the background radiation, the person region is detectable in thermal imagery. Also, shadows do not appear in thermal imagery unless the person is stationary for a long duration (shadow gradually cooling the background).

Though some classic computer vision problems are alleviated with the use of thermal imagery, common ferroelectric (chopper) sensors have their 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. Most of the previous strategies for object/person detection in thermal imagery have used "hot-spot" algorithms, relying on the assumption that the person (object) 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. Standard background-subtraction techniques alone are also rendered ineffective due to the thermal halos and polarity changes.

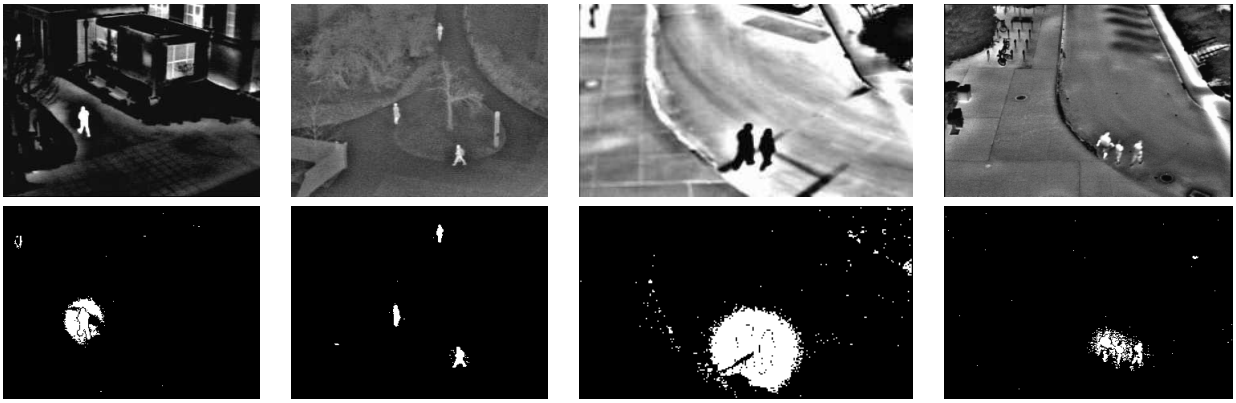
We propose a new robust background-subtraction algorithm that can be used to detect people in thermal imagery regardless of the image polarity and thermal halo. The approach does not rely on any prior shape models or motion information, and therefore the method could be particularly useful for bootstrapping more sophisticated tracking algorithms. Our approach first uses a standard background-subtraction technique to identify local regions-of-interest. The foreground and background gradient information within each region are then combined as to highlight only the person boundary. This boundary is then thinned and thresholded to form contour fragments. An A* search algorithm constrained to a local watershed segmentation is then used to complete/close any contour fragments. Finally, the contours are flood-filled to make silhouettes. We demonstrate the approach using a single set of paramters/thresholds across four very different thermal video sequences recorded from two thermal cameras.

The remainder of this paper is described as follows. We begin with a review of related work (Sect. 2). Next we describe the motivation for our approach (Sect. 3). We then describe the main components of the proposed method (Sects. 4,5,6), and present experimental results (Sect. 7). Lastly, we conclude with a summary of the research and discuss future work (Sect. 8).

2. Related Work

Several methods have been proposed for identifying people in images without background-subtraction methodologies, including the direct use of wavelets [10], coarse-to-fine edge matching [5], and motion differencing [16, 9].

*Appears in *IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, Washington DC, July 2, 2004.



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

Figure 1: Thermal imagery at different environmental conditions recorded from two different cameras and the results from statistical background-subtraction. (a) Winter-I. (b) Winter-II. (c) Summer-I. (d) Summer-II.

Most of the remaining person detection methods employ some form of background-subtraction using a single Gaussian background model [17] or a multi-modal Gaussian formulation [13]. Other approaches include the W4 method for detecting body parts and tracking [6], the three-stage (pixel/region/frame) Wallflower approach [14], a two-stage color and gradient technique [8], and a Markov chain Monte Carlo approach [19].

Recently, person detection using thermal imagery has been explored [7, 1], but these approaches rely heavily on the assumption that the person region always has a much hotter (brighter) appearance than the background (hot-spot techniques are commonly employed in thermal-based detection schemes [2, 4, 18]). We examine a new contour analysis technique for detecting people in thermal imagery that is most related to the color/gradient approach of [8].

3. Motivations for Proposed Approach

Two issues limiting the applicability of standard background-subtraction and hot-spot methods to thermal imagery produced from uncalibrated ferroelectric (chopper) thermal sensors[1] are the changing thermal polarity of objects (relative light-dark thermal intensity mapping) and the halo effect across different environmental conditions (see top row of Fig. 1). Additionally, the body of the person is not always uniformly hot or cold. Clearly, such traditional detection techniques alone will be ineffective to extract the shape of the people across the different examples shown in Fig. 1.

Two key observations regarding halos in such thermal imagery (regardless of the polarity) are that 1) thermal halos fade smoothly into the image, and 2) stronger halos cause the edge/contour information of the person within the halo to become more pronounced. Based on these observations, we propose a new background-subtraction technique for person detection that focuses on the extraction and completion of edge contours of the people within the halo regions. Because the approach relies on contours, the method is more stable and robust across very different environmental conditions, including intensity polarity switches (person can be bright or dark, or both) and different halo strengths.

[1]Next-generation microbolometers can be used to overcome some of the problems mentioned, however their resolution and quality are lower than ferroelectric sensors.

4. Region Detection

We begin by identifying localized regions-of-interest (ROIs) that contain the person (or people) and the surrounding thermal halo. We apply a standard statistical background-subtraction approach that employs a univariate Gaussian model for each pixel location (a Gaussian mixture-model [13] could also be used) and identify foreground pixels using the squared Mahalanobis distance

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

In Fig. 1, we show the background-subtraction results (with $T = 6$ and a background model for each sequence created from 30 images) for the four different examples. The thermal halos are quite prominent in Fig. 1(a),(c). Simply raising the value of T will not remove the halo in all cases. Hence, statistical background-subtraction alone is ineffective at detecting the precise shape of the person.

To extract the ROIs, we apply a 5×5 dilation operator to the background-subtracted image and employ a connected components algorithm. Any region with a size less than approximately 40 pixels is discarded.

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

Figure 2: Contour saliency. (a) ROI. (b) Foreground gradient magnitudes. (c) Foreground-background gradient-difference magnitudes. (d) CSM. (e) tCSM.

5. Contour Detection

We next examine each ROI individually to separate the person (or people) from the surrounding halo. From the earlier observations regarding thermal halos, the gradient strengths within the ROI can be used to identify much of the person boundary. For each ROI, we form a *contour saliency map* (CSM), where the value of each pixel in the CSM represents the confidence/belief of that pixel belonging to the boundary of the person.

A CSM is formed by multiplying the normalized foreground gradient magnitudes with the normalized foreground-background gradient-difference magnitudes in the ROI

$$\text{CSM} = \frac{\|\langle I_x, I_y \rangle\|}{\alpha} \times \frac{\|\langle (I_x - BG_x), (I_y - BG_y) \rangle\|}{\beta} \quad (2)$$

where the normalization factors α and β are the respective maximum magnitudes of the foreground gradients and foreground-background gradient-differences in the ROI. The range of values in the CSM is [0, 1], with larger values indicating stronger confidence of belonging to the person boundary.

The motivations for this formulation are that 1) large non-person foreground-background gradient-difference magnitudes resulting from the halo are suppressed (as they have low foreground gradient magnitudes), and 2) large non-person foreground gradient magnitudes are suppressed (as they have small foreground-background gradient-difference magnitudes). Thus, the CSM preserves the foreground gradients that are both strong *and* significantly different from the background.

For the ROI in Fig. 2(a), extracted from Fig. 1(c), we show the normalized foreground gradient magnitudes in Fig. 2(b) and the normalized foreground-background gradient-difference magnitudes in Fig. 2(c). To calculate the gradients, Gaussian derivative masks with $\sigma = .75$ were employed. We present the corresponding CSM in Fig. 2(d).

5.1. Thinning

Our next step is to produce a thinned representation of the CSM, which we call the tCSM. Since the CSM does not represent true 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. Instead, we use non-maximum suppression to thin the *foreground* gradients to create a binary mask, which is then multiplied with the CSM to produce the tCSM. The result of thinning Fig. 2(d) in this manner is shown in Fig. 2(e).

5.2. Thresholding

After thinning, we then threshold the tCSM to select the most confident segments. A simple K-Means clustering of the tCSM with 3 clusters (C_1=low, C_2=medium, C_3=high confidence values) enables an adaptive selection of the person/non-person contour pixels (motivated by the potentially multi-modal thermal intensity of pixels belonging to the person region). The pixels belonging to the lowest cluster (C_1) are discarded as the background, leaving the remaining pixels in the top two clusters (C_2, C_3) as the person boundary pixels.

5.2.1. Amplification

To improve the thresholding results we draw upon our earlier observations of thermal halos. If an object (person) has a high thermal contrast with the background, it will be surrounded by a halo that makes its boundary significantly stronger. This typically results in a distribution that is much easier to threshold. In the case of weaker thermal contrasts, the boundary strengths do not show a distinct person-background separation, and thus the tCSM is more difficult to threshold appropriately. Based on this observation, we develop a method to *amplify* the tCSM values proportional to the strength of the contrasts within the ROI.

We begin by thresholding the tCSM using the K-Means clustering approach described above to provide an initial estimate of the confidence (low, medium, high) at each pixel location in the tCSM. We then compute an amplification

factor γ from the pixel locations in the *foreground* gradient image using

$$\gamma = \min(1 - \frac{\text{median}(\text{Mag}(C_1))}{\text{median}(\text{Mag}(C_{2,3}))}, 1) \quad (3)$$

where $\text{Mag} = \|\langle I_x, I_y \rangle\|$ is the foreground gradient magnitude image. The value of γ is a measure of the gradient contrast strength within the ROI. Higher values of γ denote ROIs where the foreground gradient strengths are significantly higher than the background gradient strengths (strong halos). ROIs with more uniform gradient strengths have lower γ values (weak halos).

Using the computed amplification factor, the tCSM is adjusted as

$$\widehat{\text{tCSM}} = \text{tCSM}^{\gamma} \quad (4)$$

If γ is close to 1 the tCSM does not change significantly. However, in weak halo ROIs, lower γ values can dramatically increase the strength of the tCSM pixel values. To select the final contour pixels, we apply the K-Means clustering method (using 3 clusters) to the $\widehat{\text{tCSM}}$ and remove any pixels belonging to the bottom cluster. Other methods of thresholding, such as hysteresis, could also be applied at this stage.

We show the results using the computed amplification factor for a strong halo ROI and a weak halo ROI in Fig. 3. In Fig. 3(c-d), we show the original tCSMs. In Fig. 3(e-f), we show the amplified tCSMs ($\widehat{\text{tCSM}}$s). In the strong halo case, there was little amplification ($\gamma = .92$). In the weak halo example, the amplification was much stronger ($\gamma = .44$). The final thresholded version Fig. 3(e) is shown in Fig. 4(a).

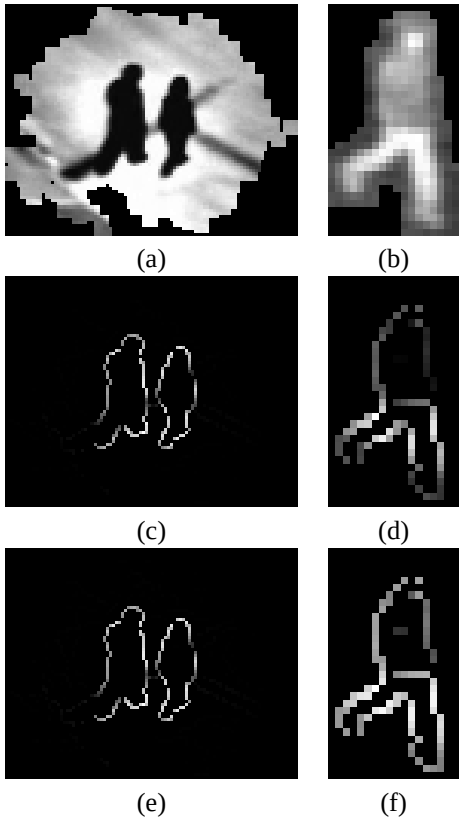


Figure 3: Amplification results. (a) Strong halo ROI. (b) Weak halo ROI. (c) Original tCSM for (a). (d) Original tCSM for (b). (e) Amplified tCSM for (a). (f) Amplified tCSM for (b).

6. Contour Completion/Closing

If the resulting thresholded $\widehat{\text{tCSM}}$ is guaranteed to have unbroken contours around the person (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). Our approach is to first complete any gaps by searching outward from each gap endpoint to find another contour pixel. Next, we verify that the figure contours 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 of the original CSM.

6.1. Watershed Segmentation

The watershed transform (WT) is a powerful mathematical morphology tool for segmenting images by partitioning image regions with watershed lines [3, 15]. When computing a WT, the image is considered as a topological relief where the elevation is proportional to the graylevel of the pixels. The determination of the watershed lines from this elevation surface can be described in terms of both topology [3] and immersion simulations [15]. 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" [3]. When the WT is applied to a gradient magnitude image, the resulting watershed lines are found along the edge ridges.

Given a gradient image, there is a high degree of overlap between its watershed lines and the result after non-maximum suppression. Hence, we can use the WT of the original CSM to provide a meaningful guide to complete any broken contours derived from the thresholded $\widehat{\text{tCSM}}$. As long as the CSM retains *some* information of the object/person boundaries, the WT will produce lines along the boundaries. Even if the WT result is highly over-segmented, the number of potential watershed lines between gaps will be smaller than all possible paths. Due to the small size of

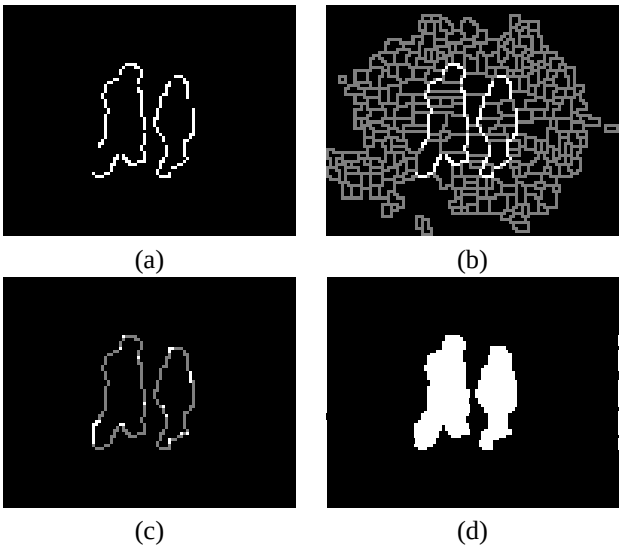


Figure 4: Watershed analysis. (a) Thresholded $\widehat{\text{tCSM}}$. (b) CSM watershed lines overlayed with (a). (c) Contour completion/closing result. (d) Flood-filled silhouettes.

the ROIs, the cost of the WT is fairly minimal. In Fig. 4(b), we show the thresholded $\widehat{\text{tCSM}}$ (from Fig. 4(a)) overlayed on the WT of the corresponding CSM (from Fig. 2(d)).

6.2. Contour Completion

We first attempt to complete any contour gaps using the watershed lines as plausible connection pathways. Each contour fragment endpoint (found using 3×3 neighborhood analysis) is forced to grow outward towards any other endpoints within a local search window (size limited to half of the vertical span of the selected contour pixels in the ROI). If no other local endpoints exist, the remaining contour points in the window are treated as target points. To force the path to grow outwards, we ignore other selected contour points within the 3×3 neighborhood of the starting endpoint.

To find the optimal path, we employ the A* search algorithm [12] that minimizes the expected cost *through* the current pixel location to reach a target point. The Euclidean distance from the current location to the closest target point is employed as the heuristic cost function. The valid paths are restricted to only the watershed lines. The outward growth of an endpoint (to a target pixel) is terminated if its path intersects any other existing contour point.

The resulting path for an endpoint may not always produce a reasonable completion of the gap. For example, there may be cases when the endpoint simply grows out and around to a pixel sitting on the endpoint's own contour segment (forming a small loop). Therefore, we incorporate a simple verification step that ensures reasonable path completion. We first find the path P_1 from the endpoint to a target point using the approach described above. Then we attempt a new path P_2 from the endpoint to the destination point, this time without ignoring the contour points around the 3×3 neighborhood of the endpoint (as is done to find P_1). If the two paths are equivalent, we keep P_1. If the paths are different, we choose the path with the maximum overlap with the unthresholded $\widehat{\text{tCSM}}$, thus choosing the path with the most "support". We compute the amount of support for a path by counting the number of pixels n in the unthresholded $\widehat{\text{tCSM}}$ that exist along that path. We then choose between P_1 and P_2 using

$$\text{P} = \begin{cases} P_1 & \frac{|n_1|}{|P_1|} > \frac{|n_2|}{|P_2|} \\ P_2 & \text{otherwise} \end{cases} \quad (5)$$

Each gap completion search uses only the original contour points (from the threshold $\widehat{\text{tCSM}}$) so that the order of the gap completion does not influence the final result. Additionally, 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.

6.3. Contour Closing

In the second stage, we ensure that every contour in the image is part of a closed loop (required for flood-filling). First, we region-grow along all contours to identify any contours that do not form a closed loop (e.g., a line connecting two closed circles is itself not closed).

For each non-closed contour, we perform the A* search strategy from one contour endpoint to the other endpoint, moving along watershed pixels not on that contour[2]. 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). If no possible path exists between the endpoints (no watershed path), we default to a direct straight-line closing between the endpoints.

The result for the thresholded $\widehat{\text{tCSM}}$ in Fig. 4(a) after completion and closing is shown in Fig. 4(c). A simple flood-fill operation can then be employed to create silhouettes (see Fig. 4(d)).

[2]For an un-closed contour with multiple endpoints (e.g., a three-prong connected contour), we compute a priority matrix [11] to select which two endpoints should be closed first, and then re-estimate the remaining unclosed contours.

7. Experiments

We examined the proposed approach on four thermal sequences recorded at very different environmental conditions, including different seasons (Winter/Summer) and time-of-day (afternoon/evening). The sequences were captured using two different Raytheon ferroelectric thermal sensor cores (300D, 250D). The number of frames in each sequence is Winter-I:1466, Winter-II:900, Summer-I:314, and Summer-II:297. Each sequence had an additional 30-frame background sequence for learning the statistical background model to identify the ROIs.

To demonstrate the generality and applicability of the approach, we extracted silhouettes from each sequence using the proposed method with the **same parameter/threshold settings for all sequences**. To give flexibility to a human operator to set the sensitivity of detection, we weight each resulting silhouette in the image with a contrast measurement calculated from the ratio of the maximum foreground-background intensity difference within the silhouette region to the full intensity range of the background image. A final sensitivity threshold could easily be used to remove any minimal-contrast (noise) regions.

In Fig. 5, we show selected frames from the sequences and the resulting (weighted) silhouettes extracted using our approach. The silhouette images demonstrate the ability of the algorithm to reliably detect people across very different thermal imagery using the same threshold settings. Furthermore, the results show that the method can *separate* multiple people contained within a single ROI.

In Fig. 5(a), the small regions in the top left corner of each image are people partially occluded by tree branches. In Fig. 5(b), different people have different thermal contrasts with the background, but all are consistently detected. Also, the silhouette shapes closely match the true person shapes in the images. In spite of the thermal similarity of the cross-walk and people in Fig. 5(c), the silhouettes were extracted and separated quite well. In Fig. 5(d), the algorithm was able to detect reasonable portions of the people despite the very low thermal person-background differences (and low gradients). Additionally, a small animal was detected moving down the stairs in the top-right corner in the first four images.

The overall results of the approach were encouraging and provide better silhouettes than could be attained using a single background-subtraction or hot-spot approach. However, there were some problems that deserve mentioning. In Fig. 6(a), the thermal intensity of the people is similar to the background cross-walk line on the pavement. This causes a reduction of the contour saliency at the overlapping pixels and therefore sometimes resulted in a contour completion growing slightly into the similar background region. Similarly, in Fig. 6(b) the thermal similarity of the people with the background caused portions of the bodies to be deleted from the silhouettes. These problems can be expected with any "intensity-based" (thermal/grayscale) method when foreground and background pixels are of similar intensity.

8. Summary

We presented a new background-subtraction method to detect people in thermal imagery over a wide range of environmental conditions (including day/night and Winter/Summer scenarios). Our approach is designed to handle problems related to halo artifacts and uncalibrated polarity switches that are typically associated with common ferroelectric (chopper) sensors. These problems render classic background-subtraction and hot-spot detection methods ineffective by themselves.

Our approach first uses a statistical background-subtraction technique to identify local regions-of-interest containing the person (or people) and the surrounding halo. The foreground and background gradient information within each region are then combined into a contour saliency map. The contour saliency map is thinned and amplified based on the strength of the thermal contrasts (halo). The most salient contours are then selected using an adaptive K-Means threshold. To complete/close any broken contour fragments, a watershed-constrained A* search strategy is used. Lastly, the contours are flood-filled to produce silhouettes.

Experiments with four thermal video sequences recorded at very different environmental conditions and a single set of parameters/thresholds for the method showed promising results, including the separation of multiple people within a single ROI. To further improve the results, we will incorporate a multi-modal background model, include additive motion information into the saliency map, and employ shaped-based models for tracking.

As the approach is not limited to only extracting silhouettes of people, we will also examine the method for detecting other objects of interest (vehicles). We additionally plan to formulate a method to quantitatively compare various background-subtraction algorithms to validate our approach.

Acknowledgements

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.

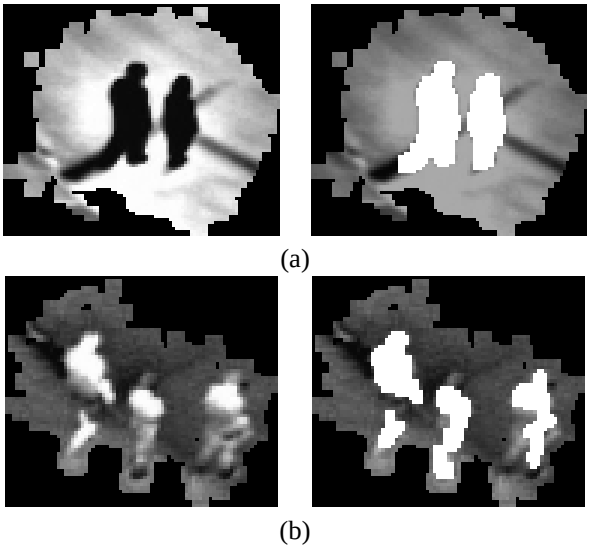


Figure 6: Problem images. (a) Silhouette extension into the background. (b) Deletion of body regions.

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

[3] M. Couprie and G. Bertrand. Topological grayscale watershed transformation. In *Vision Geometry V*, volume 3168, pages 136–146. SPIE, 1997.

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

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

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

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

[8] 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.

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

[10] 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.

[11] K. Rangarajan and M. Shah. Establishing motion correspondence. *CVGIP: Image Understanding*, 54(1):56–73, 1991.

[12] S. Russell and P. Norvig, editors. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2003.

[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] 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.

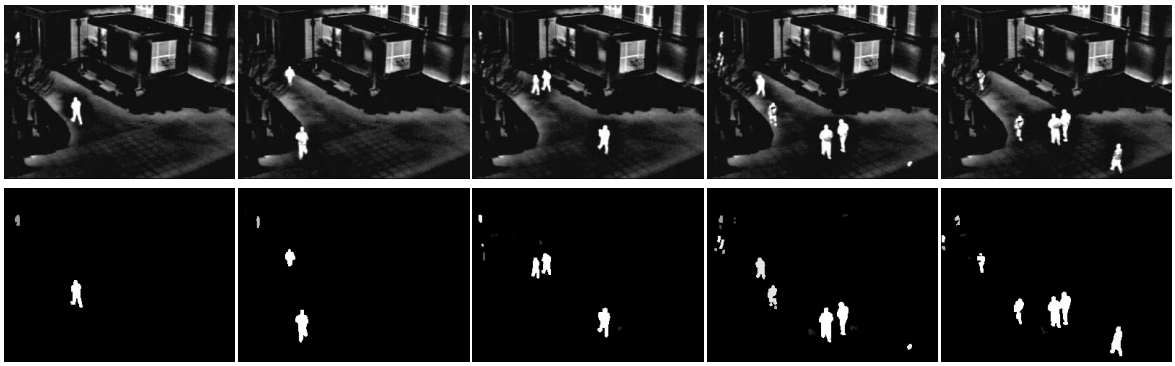
[15] L. Vincent and P. Soille. Watershed in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 13(6):583–598, 1991.

[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.

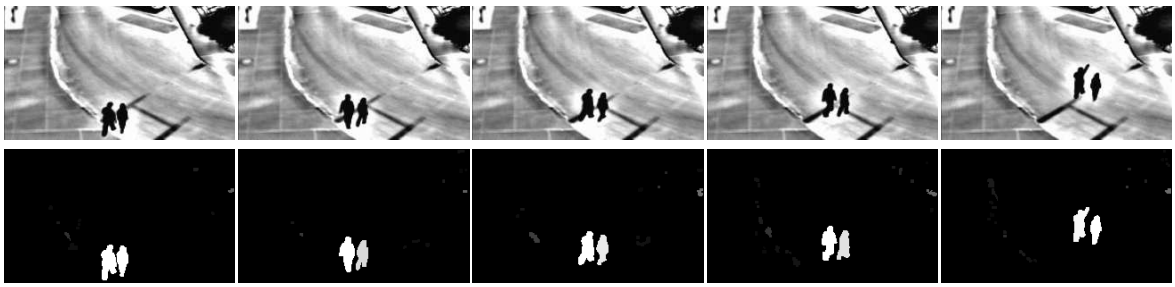
[18] A. Yilmaz, K. Shafique, and M. Shah. Target tracking in airborne forward looking infrared imagery. *Image and Vision Comp.*, 21(7):623–635, 2003.

[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.



(a) Winter-I

(b) Winter-II



(c) Summer-I

(d) Summer-II

Figure 5: Example thermal images and resulting silhouettes (contrast weighted).