# A Two-Stage Template Approach to Person Detection in Thermal Imagery*

James W. Davis  Mark A. Keck
Ohio State University
Columbus OH 43210 USA
{jwdavis,keck}@cse.ohio-state.edu

## Abstract

*We present a two-stage template-based method to detect people in widely varying thermal imagery. The approach initially performs a fast screening procedure using a generalized template to locate potential person locations. Next an AdaBoosted ensemble classifier using automatically tuned filters is employed to test the hypothesized person locations. We demonstrate and evaluate the approach using a challenging dataset of thermal imagery.*

## 1. Introduction

Automatic video surveillance systems will be expected to detect, track, and recognize human activity in a persistent 24/7 manner. Thermal video cameras offer an obvious advantage to nighttime surveillance (as shown by their widespread military and law enforcement use), but they are also applicable to daytime monitoring. In either situation, when the thermal properties of a person are different from the background (typically the case), the person regions can be detected in the video. Furthermore, traditional computer vision problems associated with shadows are minimized.

However, common ferroelectric thermal sensors have their own unique challenges, including a low SNR, white-black/hot-cold polarity changes, and halos that appear around very hot or cold objects. 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, which make standard background-subtraction and template matching techniques ineffective by themselves to detect the precise locations and shapes of the people.

In this paper we present a two-stage approach to detect people in thermal imagery that combines a special background-subtraction method [1] with an AdaBoosted template classification technique similar to that of [7]. The method first performs a fast correlation-based screening procedure to hypothesize the person locations in the image.

Figure 1: Thermal images showing large intensity variation.

To enhance the detection rate, a thermal-based background-subtraction technique [1] is employed in the screening process. The candidate regions are then examined more fully by an AdaBoosted ensemble classifier using a set of filters/classifiers adaptively modeled from the training data rather than selected from a predefined filter bank (as in [7]). The method is cast into a multi-resolution framework to detect people of different sizes and distances to the camera. We demonstrate the approach on a difficult dataset of thermal imagery (as shown in Fig. 1).

## 2. Previous Work

Several methods have been proposed for identifying people in color/grayscale images. Some examples include the direct use of wavelet features with support vector machines [6], coarse-to-fine edge template matching [3], motion/intensity AdaBoosted classifiers [7], and the size/shape of image differencing regions [4]. Several other related methods using color, texture, and stereo have also been proposed. Our approach is most closely related to the AdaBoost framework of [7], though our approach automatically adapts the filters during the AdaBoosting process.

Other person detection approaches using thermal imagery have also been proposed (e.g., [8, 5]), however most

*Appears in *Workshop on Applications of Computer Vision*, Breckenridge, Co, January 5-7, 2005.

of these methods rely heavily on the assumption that the body region is significantly hotter/brighter than the background. As shown in Fig. 1, such hot-spot techniques are not generally applicable. Our initial screening method employs the Contour Saliency Map representation [1] to robustly accommodate problematic thermal polarity switches and halos for detecting the potential person locations.

# 3. Stage-1: Screening

The approach begins with a fast screening procedure in an attempt to hypothesize only the locations of the people in the image. We create a generic person template that is correlated across the image looking for matches to take advantage of efficient software/hardware implementations for correlation-based matching. Any window location in the image that passes this initial stage is forwarded to the AdaBoost verification stage to validate the presence of a person. This approach is similar in concept to a two-stage cascade architecture [7].

As the person pixels in thermal imagery can vary considerably (as shown in Fig. 1), a simple appearance template of the pixel graylevels will not suffice. We instead use more invariant edge/gradient information and adopt the pre-processing approach of [1] to suppress the background gradient information while highlighting only the foreground object (person) edges.

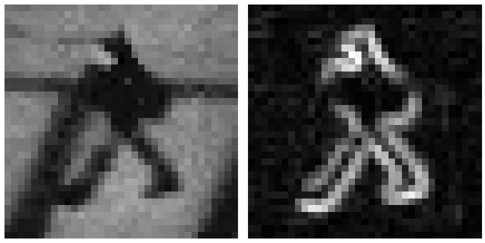
## 3.1. Contour Saliency Map

A *Contour Saliency Map* (CSM) [1] of a thermal image represents the confidence/belief of each pixel belonging to an edge contour of a foreground object (person). Initially, a background thermal image $B$ is computed (e.g., mean or median). Next, for each pixel in the input thermal image $I$, we choose the minimum of the input gradient magnitude and the input-background gradient-difference magnitude

$$\text{CSM} = \min\left(\|\langle I_x, I_y\rangle\|, \|\langle (I_x - B_x), (I_y - B_y)\rangle\|\right) \quad (1)$$

The gradient images can be computed using standard Sobel derivative masks. In [1], the two magnitude images were instead normalized and multiplied to form the CSM, but we found that the $\min$ operator produced better saliency maps.

The motivations behind the CSM representation are 1) large input-background gradient-difference magnitudes resulting from unwanted thermal halos are suppressed (as they have low input gradient magnitudes), and 2) large non-person/object input gradient magnitudes are suppressed (as they have small input-background gradient-difference magnitudes). Thus, the CSM preserves those input gradients that are both strong *and* significantly different from the background. An example CSM for a cropped thermal image containing a person and crosswalk is shown in Fig. 2.



(a) (b)

Figure 2: CSM representation. (a) Thermal image region. (b) Corresponding CSM highlighting only the person edges.

## 3.2. Generalized CSM Template

To create a generalized template for screening, we manually extract, normalize, and average several cropped windows of people from the CSM-transformed thermal images. A fixed-aspect window is manually centered around each person at any pose or orientation. To accommodate a fixed-size CSM template for differently-sized people and various person-camera distances, we construct an image pyramid for each CSM and select a cropped CSM region from the level having the best/tightest fit to the manually-selected person window. The screening threshold $T_s$ is set to the minimum (lowest) correlation value produced from the averaged template applied to the cropped CSM training examples.

## 3.3. Multi-Resolution Screening

To perform screening, each new input image is transformed into an image pyramid from which a multi-level CSM is computed. The generalized CSM template is then correlated with each level of the CSM pyramid. Any image window at a particular level with a correlation value above the template threshold $T_s$ is tagged as a potential person region.

To reduce the multiple windows that typically appear around a person, we first apply a simple 3×3 $localmax$ filter (selects a pixel only if it has the maximum value in the window) across the correlation image (at each level) to find only peaks in the detection process. Furthermore, at each level, we remove any window that overlaps (by at least 30%) any other window having a higher correlation value. The results typically the reduce the number of windows in each level to only a single candidate per person.

# 4. Stage-2: AdaBoost Classification with Adaptive Filters

The typical candidates produced by the CSM screening procedure include windows containing people, partial-person regions, and other non-person foreground objects (e.g., vehicles, animals, etc.). The task for our Stage-2 classifier is to better separate the best person matches from the remaining candidates. The basis of the approach is built upon the popular AdaBoosting learning technique [2] that was recently demonstrated for pedestrian detection in [7].

### 4.1. AdaBoost Technique

"Boosting" refers to a general method of improving the accuracy of a learning algorithm. An initial weak classifier (with accuracy only slightly better than chance) is selected. Then additional weak classifiers are added in turn to form a combined (ensemble) classifier. The technique is advantageous in that the accuracy of the ensemble classifier can be made arbitrarily high by adding additional weak classifiers until some error rate is achieved.

In "adaptive boosting" [2], referred to as AdaBoosting, a subset of the most relevant training data is used for training each additional classifier. If an example is accurately classified from the initial classifier, then its influence in the second classifier is reduced (otherwise it is increased). As an example gets correctly classified across additional classifiers, its impact in the next classifier is further reduced. With this approach, the addition of more classifiers is used to "focus in" on those examples that are most difficult to classify.

The final ensemble classification for a test example is computed from the weighted sum of the individual classifier outputs, where the weight factor for each classifier is based on its error rate on the training data.

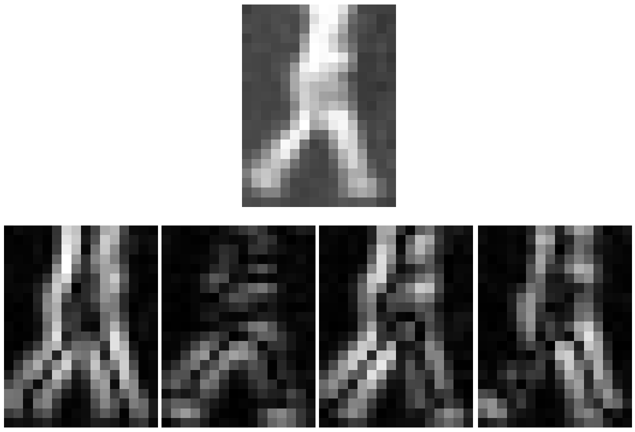
### 4.2. Adaptive Filter Selection

In [7], the filter bank, from which the best sequential classifiers were selected, was based on an *a priori* set of simple rectangular filters (at multiple scales and positions) that were applied to both the intensity image and multiple motion-difference image pairs.

In our method, we instead use the influence weights computed in the AdaBoost technique to create adaptive "holistic" templates from which small windows can be selected to create the filters. The motivation for the approach is that the filters can be adaptively tuned, rather than choosing from a fixed filter bank.

We begin with the creation of 4 feature images for each positive/negative training example using the magnitude response of 3×3 Sobel gradient operators for 0°, 90°, 45°, and 135° orientations. The 4 feature images are then normalized by the maximum value in the set. The feature images for a training example of a person is shown in Fig. 3.

For the current classifier in the AdaBoost training procedure, we use the influence weights $w(i)$ assigned by AdaBoost to each training example $T(i)$ and perform a weighted average of the feature images for each class. For



$0/180°(T^1)$ $90/270°(T^2)$ $45/225°(T^3)$ $135/315°(T^4)$

Figure 3: Original thermal image and four directional feature images using gradient operators.

feature image $k$ ($1 \leq k \leq 4$), we compute the weighted person and non-person feature images ($T_p^k, T_{np}^k$) using

$$T_p^k = \frac{\sum_{i=1}^{N_T} w(i) \cdot T^k(i) \cdot L(i)}{\sum_{i=1}^{N_T} w(i) \cdot L(i)} \quad (2)$$

$$T_{np}^k = \frac{\sum_{i=1}^{N_T} w(i) \cdot T^k(i) \cdot (1 - L(i))}{\sum_{i=1}^{N_T} w(i) \cdot (1 - L(i))} \quad (3)$$

where $L(i)$ are the binary labels assigned to the training data (0=non-person, 1=person).

The final adapted template (accounting for both positive and negative examples) for feature image $k$ is given by

$$F^k = \max\left(T_p^k - T_{np}^k, 0\right) \quad (4)$$

where larger pixel values in $F^k$ indicate locations having a stronger gradient presence from people.

The optimal filter for the current classifier is selected by finding a subregion in one of the 4 adaptive feature images $F^k$ that gives the lowest weighted error rate when applied as a filter to the training feature images. Various sizes, aspect ratios, and positions within each $F^k$ are examined to derive the optimal filters. Since we use the AdaBoost weights to generate the adaptive templates, the resulting filter in each round of AdaBoost focuses on the most difficult examples.

## 5. Experiments

### 5.1. Dataset

We collected a challenging dataset of thermal imagery to test the proposed approach. Several 360×240 thermal images of a university campus walkway intersection and street were captured over several days (morning and afternoon) using a Raytheon 300D thermal sensor core with 75mm lens mounted on an 8-story building. Example images are shown in Fig. 1. A total of 10 capture sessions were collected, with a total of 284 frames having an average of 3–4 people/frame. Three of the sessions were captured under rainy weather conditions (including people carrying umbrellas). The selected images are non-uniformly sampled at a rate much less than 30Hz and therefore object motion information is not available.

## 5.2. Training

A total of 984 people were manually marked in the images with tight-fitting bounding windows (having a fixed aspect ratio of $.76$). The smallest window (person) found in the dataset was $21 \times 16$ pixels, and was used to set the size of the screening template (similar to the template size in [7]). In our experiments we used a 3-level pyramid to accommodate the largest and smallest window sizes of the people. We selected approximately 50% of the people for training the system.

To generate the CSM for each image, we employed the median background derived from each capture session. The $21 \times 16$ CSM screening template was formed by averaging the normalized CSM templates extracted from the CSM pyramid using the manually-selected person windows. The generalized template computed from the training images (and their horizontally reflected versions) is shown in Fig. 4 and is reminiscent of the Wavelet template in [6].

For training the AdaBoosted ensemble classifier, we collected a set of negative examples of windows which passed the screening stage (prior to the $localmax$ reduction process) but did not overlap a person window by more than 30%. For each positive example in an image, we selected one of the negative examples randomly across the pyramid levels. For each of these positive and negative examples (including their reflected versions), we then computed the 4 directional gradient magnitude feature images.

Using the adaptive filter approach, AdaBoost training required 9 filters/classifiers to achieve a 100% classification for the 992 positive and 982 negative training examples. The total number of filters to choose from was 16,072 (72 possible window sizes at multiple positions within the 4 adaptive feature images). The selected filters in their correct position (and their corresponding feature images) are shown in Fig. 5.

## 5.3. Detection Results

To test the trained system, we applied the screening procedure at each pyramid level of the input images and then classified each hypothesized window using the AdaBoosted ensemble classifier.

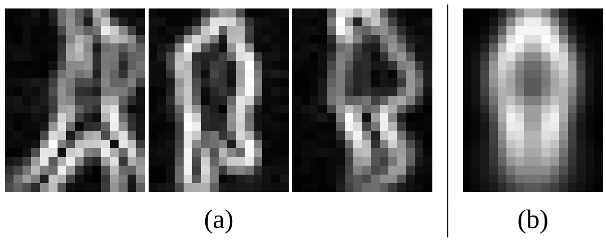


Figure 4: Screening template. (a) Example CSM training images and (b) generalized CSM screening template.

From the application of the screening template to the dataset, we generally received multiple hits per image with an average of 14 per frame over all levels, which is greatly reduced from the total number of possible windows across all three pyramid levels (140,220). Example images showing detections from the screening process are shown in Fig. 6. Lastly, we apply the previous window-reduction procedure to remove any window passing the AdaBoost ensemble classifier that overlaps (by at least 30%) any other passing window (from any level) having a higher AdaBoost confidence (unthresholded ensemble classifier value).

Some detection results after running the complete two-stage approach where the method was able to detect every person with no false positives (FP) are shown in Fig. 7. Notice that there are many cases when people are close together. There are also some images where the person is hardly noticeable from the background. For the entire dataset with a 3-level pyramid, the average center displacement of the detected person windows from the corresponding manually-marked windows was 2.09 pixels ($\pm 1.49$ SD), and the average corner error was 3.35 pixels ($\pm 2.09$ SD). These errors could potentially be further reduced with the addition of more pyramid levels.

There were also some problematic images. In Fig. 8(a), we show a person in the lower-left that was detected by two separate windows instead of one centered window. A vehicle appeared in Fig. 8(b) which was not eliminated by the AdaBoosted ensemble classifier. As there were only two images in the entire dataset containing portions of moving vehicles, a car was a rare event. With more training examples of such events, this problem could be alleviated. In Fig. 8(c), only one of the two people in the upper-right were detected due to the window overlap constraints. Groups of people are very difficult to handle with template-based detection approaches.

In [7], they report about 1 FP every 2 frames for two test sequences using their motion approach. For our dataset, we achieve about 1 FP every 47 frames. As these types of errors are obviously biased to the average number of people per frame, we additionally report the detection details along with the *Sensitivity* and *Positive Predictive Value* (PPV) in

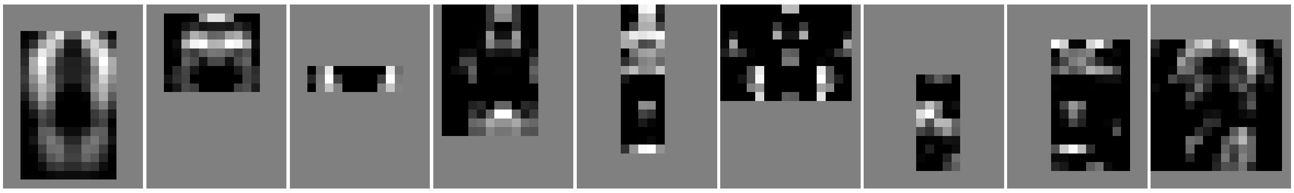



Figure 5: Resulting AdaBoosted filters (in order). The corresponding feature image for each filter is given.


Figure 6: Screening result at level-2. (a) Detections before the $localmax$ operation. (b) Final screening results.

Table 1. The Sensitivity reports the fraction of people that were correctly identified by the system, where a high Sensitivity value corresponds to a high detection rate of people, but does not account for the number of false positives. The PPV reports the fraction of detections that actually are people, where a high PPV corresponds to a low number of false positives.

The results show fairly high Sensitivity and PPV measurements for such a challenging dataset. The Sensitivity results were slightly skewed to a lower value by detecting actual people that were not manually selected because they were not fully in the scene. Furthermore, in collection 8, there were 2 people standing fairly still throughout the frames, and thus they were not often detected with the CSM screening template (lowering the Sensitivity value).

## 6. Summary and Conclusions

We presented a two-stage method to detect people in thermal imagery using a thermal-based background-suppression technique and two template-based detection methods. The initial screening stage uses a generalized person template derived from Contour Saliency Maps to quickly detect person regions while ignoring most of the background. The hypothesized person regions are then validated with an AdaBoosted ensemble classifier. Rather than selecting from a predefined set of filters to train the classifiers, our approach adaptively creates the filters from competitive gradient information of positive/negative examples. The resulting set of classifications is then reduced in an attempt to provide only a single detection per person.

The approach was demonstrated with a difficult dataset of thermal imagery with widely-varying background and person intensities. Results show that a fairly high Sensitivity and Positive Predictive Value can be achieved with the approach. We also note that the entire approach is well-suited to a parallel implementation.

In future work, we plan on extending the dataset to include additional situations involving many more distractors moving through the scene. We also will be examining other methods to handle groups of people. We additionally will seek a related CSM method to accommodate a moving camera. Lastly, we plan to incorporate color video with thermal to develop a robust fusion-based detection approach.

## References

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

[2] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 55(1):119–139, 1997.

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

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

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

[6] 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.

[7] 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.

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

| | Collection | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **1–10** |
| **# People** | 91 | 100 | 101 | 109 | 101 | 97 | 94 | 99 | 95 | 97 | 984 |
| **# TP** | 88 | 94 | 101 | 107 | 90 | 93 | 92 | 75 | 95 | 95 | 930 |
| **# FP** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 3 | 6 |
| **Sensitivity** | .97 | .94 | 1.0 | .98 | .89 | .96 | .98 | .76 | 1.0 | .98 | .95 |
| **PPV** | 1.0 | 1.0 | .99 | .99 | 1.0 | 1.0 | 1.0 | .99 | 1.0 | .97 | .99 |

Table 1: Recognition results for thermal dataset.

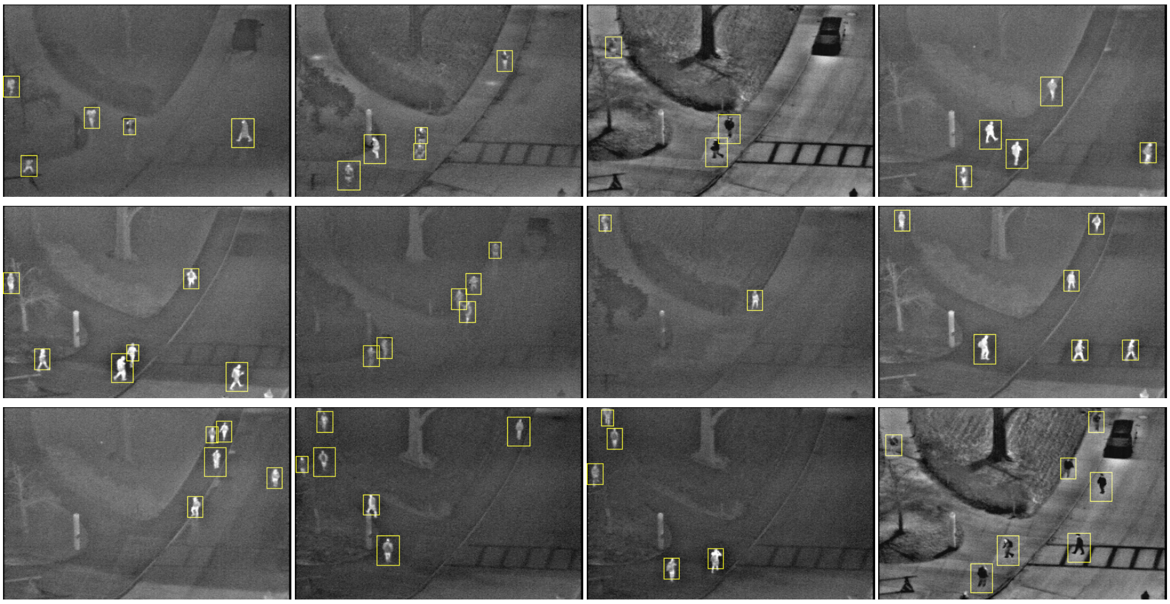


Figure 7: Example detection results with no false positives or false negatives.

(a) (b) (c)

Figure 8: Problematic images causing failure to detect people or giving false positives. (a) Single-person split. (b) Vehicle. (c) Multi-person merge.