Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

Improved Human Detection Using Image Fusion

E. Thomas Gilmore III*, Preston D. Frazier+, M. F. Chouikha*
*Department of Electrical and Computer Engineering Howard University, Washington, DC 20059, USA
+General Dynamics Corporation, Hanover, MD 21076, USA

***Abstract*— Image fusion is fundamental to several modern day image processing applications. It is often a vital preprocessing procedure to many computer vision and image processing tasks which are dependent on the acquisition of imaging data via sensors, such as infrared and visible. One such task is that of human detection. In this paper, we present improvements to our shape and heat flow-based technique of detection and classification of humans in unrestricted poses with the addition of image fusion. We focus on both rural and urban environments and demonstrate the effectiveness of using image fusion as a preprocessing procedure for improved human detection and classification. Extensive simulations using MWIR images were conducted and promising results are obtained. Receiver Operating Characteristic (ROC) analysis also showed excellent performance of the SVM-based human classification.**

I. Introduction

Infrared (IR) sensors have been applied to human detection applications such as vehicle safety, night vision, and military applications. They directly detect targets with warm temperatures in an image, providing a potentially simpler and quicker solution to human detection, especially during nighttime. However, IR sensors are much more expensive compared to optical cameras with comparable resolutions, making it less affordable for many applications. IR-based human detection has been investigated by a number of groups. Most existing research on IR-based human detection is focused on pedestrian detection in an urban environment on the street or on campus to provide assistance to the drivers or for surveillance purposes, especially during the evening [1]-[8]. Compared to non-urban environments where terrain, mountain, and/or forest scenes are the main background, urban scenes usually have artificial objects in their background such as buildings and streetlights whose temperatures are generally elevated during the evening. Vehicles also generate heat that can show up as hotspots in IR images. These background noises can make IR-based human detection more complicated. On the other hand, however, pedestrians on the street are generally in simple walking or standing upright poses which are easier to model than other complicated poses such as stretching (e.g., running and bending) or hiding, which can often occur in a non-urban environment such as in the battlefield. Human detection is obviously a more challenging situation and new methods have to be introduced, especially in dealing with the unrestricted human poses. Since little research has been done for human detection in such a non-urban environment, we have analyzed many existing algorithms designed for pedestrian detection in urban environment, and experimentally evaluated them against non-urban IR images [9]. It is shown that, as expected, these existing algorithms performed especially poorly on humans in stretching or hiding poses because they rely heavily on features of standing or walking human shapes and appearances. As a result, humans with stretching poses or partial occlusions (such as behind the trees) in the IR images are mostly missed.

In this paper, we investigate the application of Image Fusion for the purpose of improving our human detection algorithm previously presented [10]. Image fusion has been investigated by many research groups and a number of algorithms have been developed [11] - [14]. The purpose of Image fusion is to integrate images of the same target or scene from multiple sensors to produce a composite image or images that will inherit most salient features from the individual images. The fused image usually has more information about the target or scene than any of the individual images used in the fusion process. The images used for fusion here are MWIR and visible. This new method of combining image fusion to the human detection algorithm represents a natural yet powerful extension from existing pedestrian detection methods.

In section II, we briefly review the heat flow and shaped-based human detection algorithm. The application and background on Image Fusion is described in section III. Section IV presents experimental results and simulations, and section V discusses conclusions, respectively.

Manuscript received January 15, 2009. This work was supported in part by the U.S. Army Robotics Collaborative Technology Alliance (RCTA) program.

Erwin Gilmore is with the Electrical Engineering Department, Howard University, Washington, DC 20059 USA (e-mail: ethomasg@gmail.com).

Mohamed Chouikha is with the Electrical Engineering Department, Howard University, Washington, DC 20059 USA (e-mail: mchouikha@howard.edu.

II. Review of IR Based Human Detection/Classification Algorithm

A. IR Spectrum for Human Detect

Based on the distribution of infrared radiation spectrum, an IR sensor can be classified as one of the following four categories according to its wavelength [12]:

- Short Wave IR (SWIR): 0.7 - 3 μm
- Mid Wave IR (MWIR): > 3 - 6 μm
- Long Wave IR (LWIR): > 6 - 15 μm
- Far IR (FIR): > 15 - 1000 μm

IR energy is emitted by all materials and objects above 0°K as thermal radiations. The upper limit of FIR occurs in a region where it is difficult to envision the output from a source as heat (peak radiation occurs at 3°K). At normal temperature, human body radiates most strongly in the IR range at about 10 μm, which apparently corresponds to the wavelength range of LWIR. As a result, LWIR, MWIR and some FIR sensors are usually used for human detection in most applications.

B. Shape-based Feature Selection

The process of human candidate selection consists mainly of three steps: first preprocessing such as histogram equalization and segmentation by thresholding the image to obtain the hotspots, then morphological operations to suppress background noises, and finally selection of human candidates using metrics such as aspect ratio constraint, local histogram filtering, and/or morphological human model matching.

Thresholding is a technique often used to separate foreground targets from background environment based on their differences in image intensity. In a simple thresholding process, a single intensity threshold is used to generate a binary image from the original image. For example, the intensity threshold can be determined using the following equation [2].

$$Threshold = \alpha I_{mean} + \beta I_{max} \qquad (1)$$

where α and β satisfy $\alpha + \beta = 1$ and represent weights assigned to the mean intensity I_{mean} and the maximum intensity I_{max} of the original image. The best threshold setting will depend on the camera settings and the ambient conditions, e.g. temperature distribution of background objects; hence it will have to be tuned to the conditions. Determination of appropriate values for the weights, however, is not a trivial task. It is usually dependent on the specific setting of the IR camera such as brightness and/or contrast. By extensively testing our IR images using different weights, it is shown that weights $\alpha = 0.4$ and $\beta = 0.6$ perform the best, as shown in Figure 1, where 456 MWIR images with forest background were used to plot the relationship between rate of correct human selection vs. average number of non-humans selected per image with different weight values and in different aspect ratio ranges.

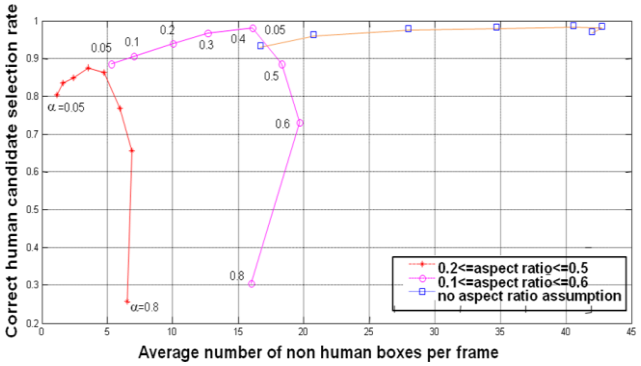
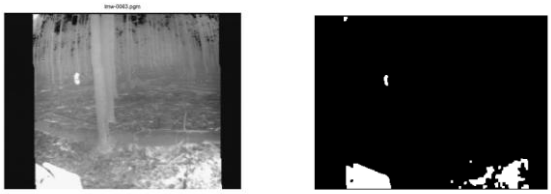
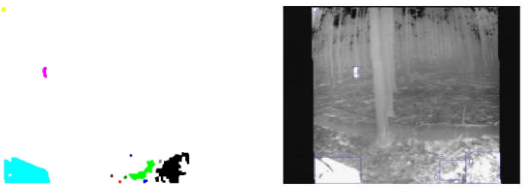


Figure 1: Threshold vs. Aspect Ratio Optimization

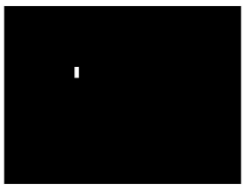
To remove isolated noises in the thresholded binary images, morphological operations of combined erosion and dilation are effectively used. Further, it is shown that local histogram of the selected human candidates can be used as a powerful filter for the elimination of false human candidates such as tree branches or electric poles [1]. This is primarily based on the fact that the intensity values of a human body in an IR image are far less uniform compared to those cylinder shaped objects. For example, the middle portion of the histogram of a bounding-box for a hotspot resulting from an electric pole is often either empty (i.e., concentration on both dark (background) and bright (pole) pixels with little gray pixels between them), or narrowly concentrated (i.e., with little or no dark or bright pixels) when the pole fills up the whole bounding-box. An example of a 'spread-out' local histogram of a human candidate is shown in Figure 2(g). Figure 2 shows an example of the process of human candidate selection.



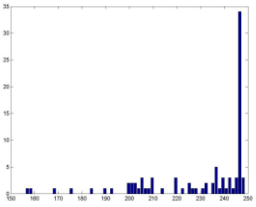
(a) Original Image (b) Thresholding/Morphological Operation



(c) Grouping of Hotspots (d) Bounding Boxes of Hotspots

(e) Applying Aspect Ratio (f) Bounding Box for Human Candidate



(g) Local Histogram of the Human Candidate

Figure 2: Example of Human Candidate Selection Process

Overall, with shape-based features we have achieved a maximum correct human candidate selection rate of 96% with a false alarm rate (or false positive rate) of around 20% in our initial experiments using the 456 MWIR images with forest background [9].

C. Heat Flow-based Feature Selection

Heat flow is a similar concept as optical flow in motion analysis using optical images [15]. Optical flow estimates motion information at pixel (x, y) at time t and $t + \delta t$ between two consecutive frames of a video camera by assuming a near-constant pixel intensity value I, which results in the following partial differential equation:

$$I_x v_x + I_y v_y = -I_t \qquad (2)$$

where (v_x, v_y) represents the motion of the pixel (x, y) or its optical flow vector.

In IR images, pixel values represent heat levels emitted by the corresponding physical points in the scene being monitored by an IR sensor, as compared to optical levels in the optical images reflected by the similar points. As a result, pixel motion in an IR image represents flow of the heat caused by motion of a warm target such as a human in the scene. We can thus use heat flow to detect relative motion of a human in an IR image.

In our method, heat flow is primarily used to locate those hotspots or bounding boxes, for reexamination, that failed to qualify the shape-based feature criteria described above. Those bounding boxes represent hotspots that were first picked up by the thresholding process, but were subsequently screened out and discarded primarily because their shape features did not fall in the range of a standing or walking human in the IR image. They were mostly treated as hotspots or noises of the background. If relative motion can be detected from those bounding boxes, however, it is strongly implicated that the targets can be human candidates in stretching poses or with partial occlusions. As a result, they will be 'rescued' from the 'trashcan' and reexamined for potential human candidates.

Relative motion of a human candidate can be detected when the magnitudes of heat flow vectors of a group of pixels inside a hotspot are larger than a threshold value determined in a similar way as that used in shape-based feature selection above. Figure 3 shows a number of examples of selection of human candidates, in both MWIR and LWIR images, in stretching poses or with partial occlusions using the proposed combined shape and heat flow method.

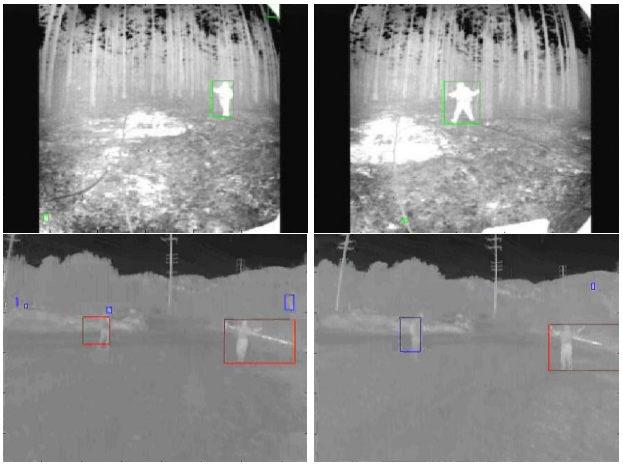


Figure 3: Example of Initial Human Candidate Selection

Preliminary experiments were performed comparing performance of initial human candidate selection using shape-based features vs. using combined shape and heat flow-based features. A total of 198 LWIR images with mountain background were used. The shape-based algorithm achieved a maximum sensitivity of 64%, but the combined shape and heat flow algorithm achieved a significantly higher maximum sensitivity of over 90% while keeping the false alarm rate at the similar level.

D. Classification

Human candidates selected above are fed to a classifier for final classification into either a human or a non-human class. We have implemented the SVM classification method [16]-[18] on our IR images, and used small templates (18x45 in size) of both gray level IR images and their edge maps as training and testing samples. A number of such training samples are shown in Figure 4.

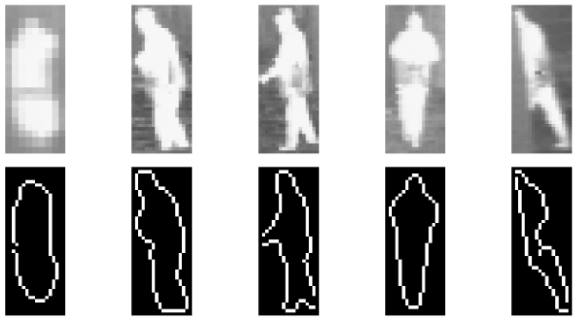


Figure 4: SVM Training Samples

III. Image Fusion Application

Multi-resolution image fusion schemes were developed to overcome the limitations of the previously introduced pixel averaging methods. The goal of these methods is to extract the salient features of each source image, e.g. edges, texture, etc., at various levels of decomposition from coarse to fine, and then aggregate them to create a fused image. The pyramid based schemes first put these concepts into practice. These methods generally produce sharp, high-contrast images that are clearly more appealing and have more information content that the simpler weighted pixel averaging techniques.

First investigated in the early 1980's, the concept of the image pyramid was used as a fast method of representing the multi-resolution information contained within an image in a manner that reflects the multiple scales of processing in the human visual system [19]. The image pyramid is basically a data structure made of a series of low-pass or band-pass copies of the image, each depicting pattern information of a different scale.

The most common example is the image pyramid, whose construction begins by convolving a source image G_0 with a Gaussian kernel K. The filtered image is then sub-sampled by selecting only every other row and column to generate a new image G_1 with half the width and height of the original image G_0. This combination of sub-sampling and convolution is known as a REDUCE operation and defined by:

$$G_1(x,y) = \sum_{u=-p}^{p}\sum_{v=-p}^{p} K(u,v)G_0(2x+u, 2y+v) \quad (3)$$

where the Gaussian kernel K is usually small, i.e. 3×3 or 5×5, for rapid execution. This process is then repeated with G_1 to develop G_2, and so on, until a pyramid of images $G_0, G_1, G_2, \ldots, G_N$ are produced. High spatial frequencies are lost when stepping from one level of the pyramid to the next due to the reduction in the resolution and sampling density. This is interpreted as a loss of salient image detail.

To compare the various image contents now available at each level of the Gaussian pyramid, the EXPAND operator is used. Basically, this consists of duplicating each row and column in the image G_{k+1} and convolving the result with the Gaussian kernel K to generate the new image E_K of the same dimensions as G_k. The EXPAND operator can be expressed by the following:

$$E_k(x,y) = \sum_{u=-p}^{p}\sum_{v=-p}^{p} K(u,v)G_{k+1}(\text{floor}[(x+u)/2], \text{floor}[(y+v)/2]) \quad (4)$$

A new image is then created from the difference between reduced image G_k and expanded image E_K:

$$L_k(x,y) = G_k(x,y) - E_k(x,y) \quad (5)$$

which captures the high frequency spatial details of the *kth* level of the Gaussian pyramid. Thus, new pyramids of varying resolutions are determined from the different Gaussian pyramids, i.e. $L_0, L_1, \ldots, L_{N-1}$, which represent salient information in the original image. This structure is known as the Laplacian Pyramid due to the Laplacian operator that is utilized and was first used for image compression applications [20], [21] and then as an image fusion scheme [22].

A. *Fusion using Laplacian Pyramid*

The Laplacian pyramid fusion method consists of an iterative process of calculating Gaussian and Laplacian pyramids of each source image, fusing the Laplacian images at each pyramid level by selecting the pixel with larger absolute values, combining the fused Laplacian pyramid with the combined pyramid expanded from the lower level, and expanding the combined pyramids to the upper level.

IV. Experiments and Simulations

We employed a Laplacian Pyramid fusion scheme to generate a series of fused images to be utilized as input to the Detection and Classification scheme. Figure 5 depicts a representative test image where the green bounding boxes indicate a correct detection and red bounding boxes indicate a correct rejection, respectively. Initial training and testing experiments were performed with sample fused images (bounding boxes) of selected human candidates (426 for training – 253 humans and 173 non-humans, and 1146 for testing – 654 humans and 492 non-humans) using the SVM classifier. All sample images (bounding boxes) were scaled to the same template size of 18x45 before being fed to the classifier.

The Receiver Operating Characteristic (ROC) curves for an SVM classifier with a quadratic kernel function were generated and are shown in Figure 6, which demonstrated significant improvement over the non-fusion based human detection scheme.

Figure 5: Sample Test Image fused via Laplacian Pyramid (Green, correct detection; Red, correct rejection)

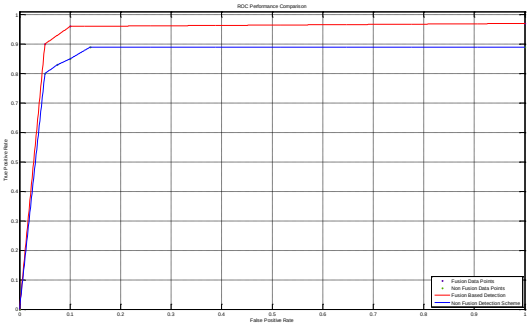


Figure 6: ROC Performance Comparison

V. Conclusion

We have developed an improvement to our original method of human detection using IR images. This method incorporated image fusion as a preprocessing task to the combined shape and heat flow-based detection scheme. Preliminary experiments using a large number of IR images have shown that this new method has achieved significant performance improvement over the original algorithm. The ROC curves also confirmed the excellent performance of the SVM-based human candidate classification.

VI. References

[1] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf, and M.M. Meinecke, "Pedestrian detection for driver assistance using multiresolution infrared vision", IEEE Trans. on Vehicular Technology, 53 (6), 2004.

[2] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision", IEEE Trans. on Intelligent Transportation Systems, 6 (1), 2005.

[3] M. Yasuno, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images", Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04), 2004.

[4] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images", IEEE Trans. on Vehicular Technology, 53 (6), 2004.

[5] X. Liu and K. Fujimura, "Pedestrian detection using stereo night vision", IEEE Trans. on Vehicular Technology, 53 (6), 2004.

[6] C. Dai, Y. Zheng, and X. Li, "Layered representation for pedestrian detection and tracking in infrared imagery", Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 3, pp. 20-26, 2005.

[7] M. Bertozzi, A. Broggi, C. Hilario Gomez, R.I. Fedriga, G. Vezzoni, and M. Del Rose, "Pedestrian detection in far infrared images based on the use of probabilistic templates", Proc. 2007 IEEE Intelligent Vehicles Symposium, 2007.

[8] Yu-Ting Chen and Chu-Song Chen, "A Cascade of Feed-Forward Classifiers for Fast Pedestrian Detection", Lecture Notes in Computer Science, Volume 4843, pp. 905-914, Springer, Berlin, November 2007.

[9] J.Zeng, A. Sayedelahl, M. Chouikha, T. Gilmore, and P. Frazier, "Human detection in non-urban environment using infrared images", Proc. Sixth International Conference on Information, Communications, and Signal Processing, Singapore, December 2007.

[10] J. Zeng, A. Sayedelah, H. Laryea, and M. Chouikha, "Enhanced Human Detection in Non-urban Using Combined Shape and Heat Flow Features," 2008 IEEE International Conference on Robotics and Automation, ICRA 2008, Pasadena, California, on May , 2008.

[11] Z. Zhang and R.S. Blum, "A Categorization of Multiscale Decomposition-based Image Fusion Schemes with a Performance Study for a Digital Camera Applicatoin," Proceeding of the IEEE, vol. 87, no. 8, pp. 1315-1326, 1999.

[12] P. Scheunders and S. DeBacker, "Multispectral Image Fusion and Merging Using Multiscale Fundamental Forms," Proc. IEEE International Conference on Image Processing, 2001.

[13] D. Rajan and S. Chaudhuri, "Data Fusion Techniques for Super-Resolution Imaging," Information Fusion, 3, pp. 25-38.

[14] L.A. Chan and Z.S. Der, and N.M. Nasrabadi, "Dualband FLIR Fusion for Automatic Target Recognition," Information Fusion, 4, pp. 35-45.

[15] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", Proc. Imaging Understanding Workshop, pp. 121-130,1981.

[16] V.N. Vapnik, The Nature of Statistical Learning Theory (2nd Ed.), Springer, New York, 1999.

[17] [10] J. C. Platt, “Fast training of support vector machines using sequential minimal optimization”, Chap. 12 in Advances in Kernel Methods - Support Vector Learning, B. Scholkopf, C. Burges, and A. J. Smola, Eds., pp 185--208, MIT Press, Cambridge, MA, 1999.

[18] N. Cristianini and J. Shawe-Taylor, Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.

[19] E. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt, and J.M. Ogden. Pyramid Methods in Image Processing. RCA Engineer 29, 33-41, 1984.

[20] P.J. Burt and E.H. Adelson. The Laplacian Pyramid as a Compact Image Code. IEEE Trans. Commun. COM-31, 532-540, 1983.

[21] P.J. Burt. The Pyramid as a Structure for Efficient Computation. Multi-Resolution Image Processing and Analysis.

[22] P.J. Burt and E.H. Adelson. Merging Images through Pattern Decomposition. Applications of Digital Image Processing VIII, Proc. SPIE 575, 173-181, 1985.