Flood-survivors detection using IR imagery on an autonomous drone

Sumant Sharma
Department of Aeronautcs and Astronautics
Stanford University
Email: sharmas@stanford.edu

***Abstract*—In the search and rescue efforts soon after disaster such as floods, the time critical activities of survivor detection and localization can be solved by using thermal long-wave infrared (LWIR) cameras which are more robust to illumination and background textures than visual cameras. This particular problem is especially challenging due to the limited computational power available on-board commercial drone platforms and the requirement of real-time detection and localization. However, the detection of humans in low resolution infrared imagery is possible due to the few hot spots that appear due to the heat signature. We propose a two-stage approach for human detection in LWIR images: (1) the application of Maximally Stable Extremal Regions (MSER) to detect hot spots instead of background subtraction or sliding window and (2) the verification of the detected hot spots using Integral Channel Features (ICF) based descriptors and a Naive Bayes classifier. The approach is validated by testing on an LWIR image dataset containing low resolution videos in real-time. The approach is novel since it achieves high detection rates while possessing a low computational runtime, and unlike several related works in human detection, without assuming that the targets are moving in the image frame.**

I. Introduction

Survivor detection and localization is an important part of search and rescue efforts immediately following disasters such as floods. Autonomous Unmanned Aerial Vehicles (UAV) in conjunction with digital image processing can assist rescue efforts by quickly scanning for large swaths of areas and alerting first responders. However, achieving a high true positive rate with low false positive and false negative rates is challenging due to the limited computational power available on-board, low resolution imagery, and changing background as well as illumination conditions. Use of long-wave infrared (LWIR) imagery is suitable for this application since humans have a distinct heat signature compared to the background and the imagery does not suffer from changing illumination conditions. In Figure 1, a side-by-side comparison of LWIR and visual imagery is shown to exhibit the robustness of LWIR imagery to shadows. The variation in the image intensities of humans in LWIR imagery across different cameras is much less compared to the variation across visual cameras. This allows us to use simpler detection algorithms which do not need to normalize the image intensities to form feature descriptors.

In this paper, I propose an approach described by Teutsch et al. [1] for detection and localization of humans in LWIR imagery captured by a camera mounted on a UAV. In contrast to other papers, this approach avoids the constraints such as the assumption of a stationary camera or known backgrounds. The approach has two stages: in the first one, Maximally Stable Extremal Regions (MSER) are used to detect hot spots and normalize the size each region to 16 x 32 pixels. In the second one, humans and clutter are classified using machine learning algorithms. In particular we use descriptors based on Integral Channel Features which are then classified by three different machine learning algorithms. Validation is done using the OTCBVS dataset acquired by stationary cameras.

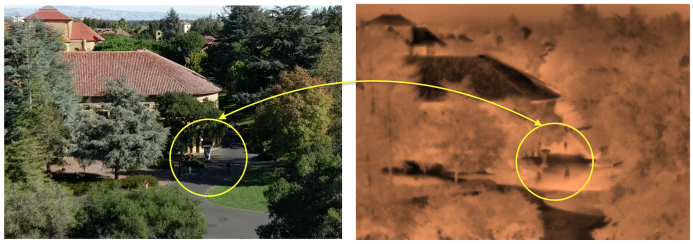

Fig. 1. Comparison of the images captured by a visual camera (left) and an LWIR camera (right) of the same scene. Note the humans in the yellow circles.

The remainder of the paper is organized as follows: literature related to human detection in LWIR images is reviewed in Section 2. Hot spot detection is described in Section 3 and classification in Section 4. Experimental results are given in Section 5. We conclude in Section 6.

A. Related Work

The approaches reviewed in context of human detection in LWIR imagery usually deal with simplifying constraints such as assuming stationary camera or detecting only moving persons [2], [3]. Regions of interest (ROIs) are detected either with background subtraction [4], keypoint detectors [5], sliding window [6], or thresholding methods such as MSER [7]. All approaches except of background subtraction can be applied with a moving camera.

These ROIs can be verified using machine learning algorithms. Davis and Keck [4] calculate gradient magnitudes in four different directions and automatically learn a weighted filter bank using AdaBoost classifier. Li et al. [8] use a combination of Histogram of Oriented Gradient (HOG) features with geometric constraints and a linear Support Vector Machine (SVM) classifier. Leykin et al. [3] model and recognize human

motion after tracking. Teutsch [7] proposes a descriptor based on Hu moments, central moments, and Haralick features followed by SVM classification. Jungling and Arens [5] use Speeded Up Robust Features (SURF) to detect and classify body parts and assemble them with an Implicit Shape Model (ISM).

II. Hot Spot Detection

As per the literature for related work, three methods are applicable for hot spot detection: keypoint detection, sliding window, and MSER. Keypoint detection is not suitable as it does not detect many features in low resolution images, leading to partial or missed detections. Sliding window approaches are also not suitable as these are computationally expensive and require searches across multiple scales in an image pyramid to account for scale changes. Thus, MSER is pursued as it has low computational requirements and is robust to low resolution imagery. As learned in class, MSERs are essentially the result of blob detection based on connected component labeling. The underlying assumption that allows us to use MSER in this problem is that the body temperature of persons in LWIR is higher than the temperature of the immediate background surrounding the person. Do note that, this will lead to false detections of cars, trees, and street lights. Moreover, depending on the homogeneity of the body temperature, partial MSER detection is possible. Additionally, adjacent persons could be detected in a single MSER.

MSERs are the result of a blob detection method based on thresholding and connected component labeling [24]. The application of MSER detection in this paper follows the assumption that the body temperature of persons in LWIR images is generally higher than the temperature of the surrounding background. This is true for many outdoor scenarios. Additional MSERs will be detected for background hot spots such as warm engines, open windows and doors, or areas heated up by the sun. Depending on the number and size of hot spots, merged detections will appear affecting the human blob shape. We use the following hierarchical MSER approach in order to handle such merged detections of either several persons or persons with background. Since the ultimate goal of the project is to locate as many survivors as possible while minimizing false positives, we do not distinguish MSERs corresponding to merged detections of multiple people. Typical hot spots detected in LWIR imagery of humans is shown in Figure 2.

III. Hot Spot Classification

The purpose of classification is to verify whether the detected hot spots are originating from humans or not. In order to achieve reliable classification while maintaining low computational requirements, appropriate feature descriptors need to be created. Following that, we can utilize machine learning algorithms to generate classifiers that can distinguish between “humans” and “not humans”.

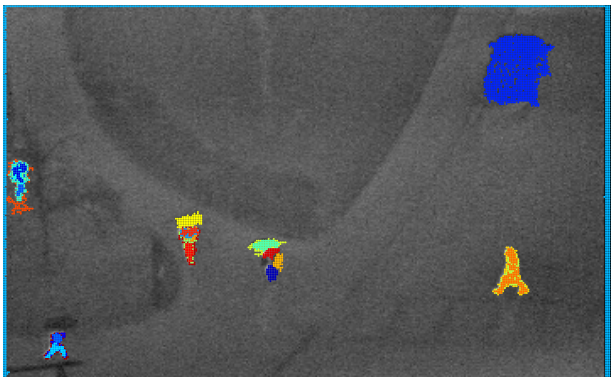

Fig. 2. Typical hot spots detected in LWIR imagery of humans.

A. Features

For search and rescue operations using UAVs, survivors can appear in a variety of different resolutions depending on their distance from the camera. We can account for these scale changes by normalizing the size of each detected hot spot. The scale of each detected hot spot is scaled to 16 x 32 pixels using bi-linear interpolation. For each hot spot, we then calculate descriptors based on Integral Channel Features [9].

These features are based on the gradient magnitudes of the image. The gradient magnitude can be obtained using the Sobel operator. For example, the gradient along the x- and y-axis can be obtained by convolving the MSER image with G_x and G_y:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \quad (1)$$

The magnitudes are calculated along six orientations (at 30 degree intervals between x- and y-axis) which capture the directional component of the gradient. Additionally, an approximation of the total gradient magnitude is calculated using

$$G = \sqrt{G_x^2 + G_y^2} \quad (2)$$

Essentially each 16 x 32 MSER yields seven 16 x 32 gradient images. Local sums of the image intensities are then calculated in randomly picked rectangular regions along all seven images and concatenated to set up the 2000 x 1 descriptor. These local sums are also know as first-order features. If needed, higher order features can be calculated by summing randomly picked local sums but they were not used in this analysis.

B. Classification

Besides the evaluation of state-of-the-art classifiers such as SVM using different kernel types, we also analyze the Naive Bayes classifier since SVMs are known to be computationally expensive and slight changes in pose can make some features not fit the model anymore leading to poor classification performance. The Naive Bayes classifier is known to be fast and provides good classification performance even when the assumption of conditional independence of the used features

is violated [10]. The Naive Bayes decision boundary is given by:

$$\text{class}_{\text{NB}}(\mathbf{f}) = \arg\max \left\{ P(c_i) \cdot \prod_{j}^{n} P(f_j|c_i) \right\} \quad (3)$$

where $f = (f_1, ...f_n)$ is the feature vector, $P(c_i)$ is the prior probability for the class c_i with $i \in 0, 1$ and $P(f_j|c_i)$ is the likelihood for the feature f_j given class c_i. The product of these likelihoods is based on the naive assumption that the features f_j of a descriptor are conditionally independent.

IV. Experimental Results

The dataset is separated in disjoint training and test sets to enable supervised learning of the classifier models. We use the dataset provided by OTCVBS [4], which contains 286 images of 360 x 240 resolution each. The dataset has 10 sequences corresponding to different videos captured at separate times. I use the first four as my training set and the last six as the test set. The ground truth contains bounding box annotations of the 2814 instances of humans present in the imagery. The MSER results calculated for the training and test datasets were labeled manually for person or background MSER. Figure 3 shows some example person and background MSERs in the four datasets. The appearance of person and background hot spots varies across and inside the datasets. Note that partial or

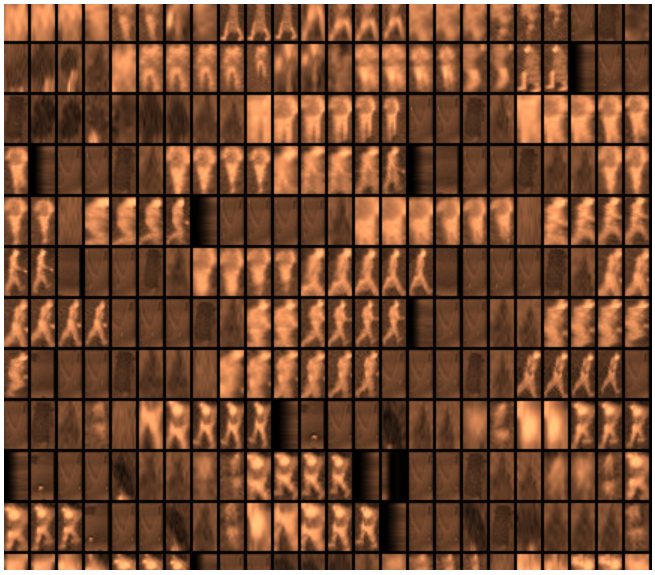

Fig. 3. A montage of some of the detected MSERs in the training set used in the experiments.

merged detections were not considered for classifier training since we are only interested in complete and distinctly detected humans in the imagery. For each MSER we calculate each of the descriptors as discussed earlier and classify it with three classifiers. We use an SVM with a linear kernel, an SVM with Radial Basis Function (RBF) kernel, and the Naive Bayes classifier. All classifier training was done using MATLAB's Statistics and Machine Learning Toolbox. Moreover, the hyperparameters `kernelScale` and `boxConstraint` were optimized to minimize five-fold cross-validation loss. In figure 4, we show the values tried for the hyperparameters and the corresponding loss. In Figure 5, the Receiver Operating

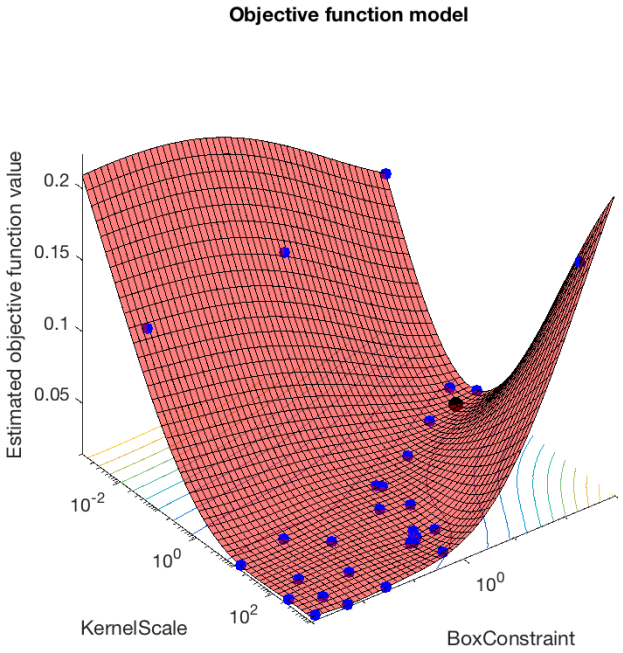

Fig. 4. Hyperparameter optimization for the liner SVM classifier.

Characteristic (ROC) curves for each classifier are calculated to exhibit the training results. The Area Under Curve (AUC) was calculated for a compact presentation of our results. The

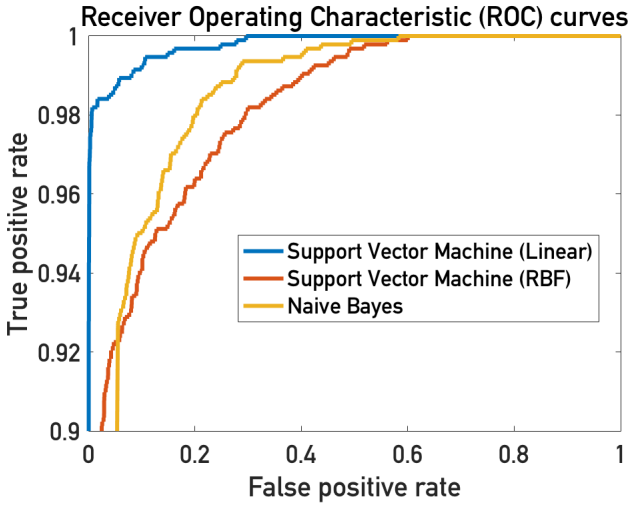

Fig. 5. The associated ROC curves for each classifier to exhibit the training results.

Linear SVM trained on the ICF features performs the best out the three approaches. As seen in the ROC curves, it correctly

classified more than 99% of the human instances with a False Positive (FP) rate of 15%. The Naive Bayes performed slightly worse with a true positive (TP) rate of 98% with a FP rate of 20%. The SVM with the RBF kernel function performed the worst with 97% TP rate and 30% FP rate.

Finally, the three classifiers were used on the test set and a few of the resulting images are shown in Figure 6. Note that were successfully able to detect the humans while avoiding the hot spots originating from the truck and the emergency pole. For a more complete real-time demonstration of the Naive Bayes + ICF approach, please see http://imgur.com/a/znUm6. The results summarizing the experiments are presented in Figure 7. We can see that the Naive Bayes classifier trained on the ICF features had the highest TP rate as well as the minimum False Negative (FN) rate. This makes this approach the most desirable for a survivor-detection scenario where detecting all possible humans is required. However, this approach does seem to have the highest FP rate. The SVM with the linear kernel trained on ICF features has the lowest FP rate. We also tested the same algorithm on an independently obtained LWIR imagery from a FLIR IR camera. A typical image of two humans outdoors overlaid with the detections is shown in Figure 8. The approach was successful in detecting the two humans in this footage, however, it also picked up a number of false positives originating from the tiled roof. Moreover, the bounding box on the human on the right is not of the correct size as the associated MSER is merging with the relative hot floor.

Fig. 6. Test set results obtained from applying the Naive Bayes classifier trained on ICF features.

V. Conclusion

In this project, I followed a two-step approach for detecting humans in real-time using LWIR imagery captured by a camera mounted on a UAV. This particular approach focuses on the low resolution imagery typical of UAV platforms and utilizes algorithms which can run on the limited computational power available. The approach is also robust to changing illumination and background conditions, unlike approaches relying on visual imagery. The approach consists of MSER hot spot detection followed by classification using Integral Channel Features and a Naive Bayes classifier. We also showed the supremacy of this approach over SVM classifiers trained on the same features using a linear kernel and the radial basis function, respectively. However, this approach is prone to be very slow in high resolution imagery as the number of MSERs detected in such images can be quite large. The approach also performs poorly when applied to images with humans standing on relatively hot surfaces as the MSERs of the humans and the ground gets merged. Future work on this problem can include ground plane detection to isolate the ground MSERs from the human MSERs. This work can also be extended to include imagery from multiple view-points while simultaneously tracking the humans to provide a very accurate and robust count of survivors.

	True Positive	False Positive	False Negative
SVM (Linear)	86.6%	3.0%	13.4%
SVM (RBF)	86.6%	6.3%	13.4%
Naive Bayes	89.9%	15.2%	10.1%

Fig. 7. Results after applying the three classifiers to the test set of LWIR imagery from the OTCVBS dataset.

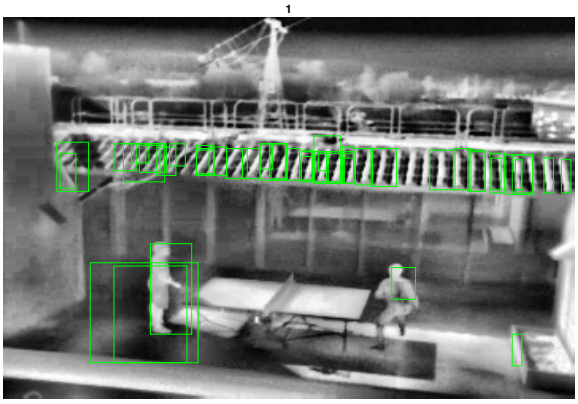

Fig. 8. Results after applying the the Naive Bayes classifier to an independently obtained imagery from a FLIR IR camera.

ACKNOWLEDGMENT

I would like to thank Professor Gordon Wetzstein and the EE368 course staff.

REFERENCES

[1] M. Teutsch, M. Thomas, M. Huber, and F. Iosb, "Low Resolution Person Detection with a Moving Thermal Infrared Camera by Hot Spot Classification," pp. 209–216, 2014.

[2] J. Wang, G. Bebis, and R. Miller, "Robust video-based surveillance by integrating target detection with tracking," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2006, 2006.

[3] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-visible video fusion for moving target tracking and pedestrian classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[4] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 162–182, 2007.

[5] K. Jüngling and M. Arens, "Feature based person detection beyond the visible spectrum," *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 30–37, 2009.

[6] W. Li, D. Zheng, T. Zhao, and M. Yang, "An effective approach to pedestrian detection in thermal imagery," *Proceedings - International Conference on Natural Computation*, no. Icnc, pp. 325–329, 2012.

[7] M. Teutsch and T. Müller, "Hot spot detection and classification in LWIR videos for person recognition," vol. 8744, p. 87440F, 2013. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2015754

[8] E. S. Jeon, J. S. Choi, J. H. Lee, K. Y. Shin, Y. G. Kim, T. T. Le, and K. R. Park, "Human detection based on the generation of a background image by using a far-infrared light camera," *Sensors (Switzerland)*, vol. 15, no. 3, pp. 6763–6788, 2015.

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," *BMVC 2009 London England*, pp. 1–11, 2009. [Online]. Available: http://www.loni.ucla.edu/$\sim$ztu/publication/dollarBMVC09ChnFtrs_0.pdf

[10] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Los," *Machine Learning*, vol. 29, pp. 103–130, 1997. [Online]. Available: http://link.springer.com/article/10.1023/A:1007413511361