MULTI-STAGE INFRARED STATIONARY HUMAN DETECTION

Alex Lipchen Chan

U.S. Army Research Laboratory, Adelphi, MD 20783, USA
alex.chan@us.army.mil

ABSTRACT

Detecting stationary human targets is crucial in ensuring safe operation of unmanned ground vehicles. In this paper, a multi-stage detection algorithm for stationary humans in infrared imagery is proposed. This algorithm first applies an efficient feature-based anomalies detection algorithm to search the entire input image, which is followed by an eigen-neural-based clutter rejecter that examines only the portions of the input image identified by the first algorithm, and culminates with a simple evidence integrator that combines the results from the two previous stages. The proposed algorithm was evaluated using a challenging set of infrared images, and the results support the usefulness of this multi-stage human detection architecture.

Index Terms—stationary human detection, clutter rejection, FLIR imagery, multilayer perceptron

1. INTRODUCTION

Empowered by the huge advances in sensor technologies, computational power, and image processing algorithms, autonomous systems are moving closer to the center stage of both military and civilian arenas in the 21st century. For example, unmanned ground vehicles (UGV) of various sizes have been developed to perform a range of activities that are deemed too hazardous or inaccessible to humans. For the larger and heavier UGV to operate safely around humans, however, it is critically important to ensure that the UGV are able to detect humans near their intended path and within their maneuvering zone. Compared to moving human targets, stationary humans are much more difficult to detect, especially when target occlusions, different poses and grouping, clothing coverage and variations, diurnal and seasonal changes, busy urban background, and other exacerbating factors are involved. Some of these difficulties may be mitigated by using forward-looking infrared (FLIR) sensors, while some researchers may choose to confine the scope of problem to non-urban environments only [1].

In this paper, a multi-stage detection algorithm is proposed to detect stationary human targets in various urban and indoor settings using FLIR imagery. This multi-stage detection algorithm consists of a set of intuitive image features and two distinct feed-forward neural networks that are cascaded together. The efficient feature extractors in the first stage quickly process the whole input scene and identify potential target areas for further scrutiny. These potential target areas are then examined by more complex but accurate stages that follow. In this way, the detection performance can be gradually improved along the stages, while the computational costs are added effectively. In addition to evaluating the overall multi-stage architecture, the intermediate results of each added stage are also discussed. Details of the proposed algorithm and its experimental results are provided in the next two sections, while a short conclusion constitutes the final section.

2. MULTI-STAGE TARGET DETECTOR

The proposed stationary human detector consists of three distinct stages. The first stage is a computationally simple cued, referred to as the Detalex detector in this paper, whose function is to estimate the locations and sizes of potential targets on the input images based on several localized image features. The selected image locations are further examined by a more complicated clutter rejecter in the second stage, which is referred to as EIGMLP in this paper because it is built with an eigenspace transformation and a multilayer perceptron (MLP). Finally, the results of Detalex and EIGMLP are nonlinearly combined in the third and final stage by an Evidence Combining MLP (ECMLP). Consequently, three different configurations or architectures of this multi-stage detector, as depicted in Figure 1, are presented and discussed in this section.

2.1. First Stage: Detalex Detector

As shown in Figure 1(a), the Detalex constitutes the first and simplest configuration of the proposed algorithm. The local features used in Detalex are computationally simple and functionally monotonic in nature. Each feature is calculated for every pixel in the image, and some knowledge of the range is assumed in order to determine approximate target size. However, these features are not highly range sensitive, because the algorithm calculates them at coarsely sampled ranges between the given minimum and maximum ranges.

In this paper, a multi-stage detection algorithm is proposed to detect stationary human targets in various urban and indoor settings using FLIR imagery. This multi-stage detection algorithm consists of a set of intuitive image features and two distinct feed-forward neural networks that are cascaded together. The efficient feature extractors in the first stage quickly process the whole input scene and identify potential target areas for further scrutiny. These potential target areas are then examined by more complex but accurate stages that follow. In this way, the detection performance can be gradually improved along the stages, while the computational costs are added effectively. In addition to evaluating the overall multi-stage architecture, the intermediate results of each added stage are also discussed. Details of the proposed algorithm and its experimental results are provided in the next two sections, while a short conclusion constitutes the final section.
strength, and local variation. Each of these features exhibits a different level of detection performance when applied to a given data set [2]. The relative importance of these features tends to change with the characteristics of the data set.

The maximum gray level, $F_{x,y}^9$, is the highest gray level within a roughly human-sized rectangle centered at the pixel $(x, y)$. It was chosen because some human targets in FLIR images may have a few pixels that are significantly hotter than the rest of the target or the background. This feature is defined as: $F_{x,y}^9 = \max_{(m,n) \in N_{in}(x,y)} f(m,n)$, where $f(m,n)$ is the gray level of the pixel in the $m$-th column and $n$-th row, and $N_{in}(x,y)$ is the neighborhood surrounding the pixel $(x,y)$ and delimited by a rectangle proportional to the expected human size in a given range.

The contrast-box feature, $F_{x,y}^1$, measures the average gray level over a human-sized region and compares it to the average gray level of the local background. It was chosen because many pixels pertaining to human targets, in addition to those extra warm pixels captured by $F^9$, are still warmer than the natural background in general. The value of this feature at pixel $(x,y)$ is computed as:

$$F_{x,y}^1 = \frac{1}{n_{in}} \sum_{m,n}^{N_{in}(x,y)} f(m,n) - \frac{1}{n_{out}} \sum_{m,n}^{N_{out}(x,y)} f(m,n),$$

where $N_{in}(x,y)$ or the inner box is the same human-sized neighborhood previously defined for feature $F^9$, while the perimmetrical neighborhood $N_{out}(x,y)$ or the outer box is a larger rectangle centered at pixel $(x,y)$, which engulfs but excludes all pixels within the inner box. Finally, the $n_{in}$ and $n_{out}$ is the number of pixels in $N_{in}$ and $N_{out}$, respectively.

The gradient strength feature, $F_{x,y}^2$, was chosen because human appearance tends to show sharper internal detail than natural objects, even when the average intensity is similar. To prevent large regions of background that exhibit higher than normal variation from exerting a high value for this feature, the average gradient strength of the local background (outer box) is subtracted from the average gradient strength of the human-sized region (inner box). For pixel $(x,y)$, this feature is calculated as:

$$F_{x,y}^2 = \frac{1}{n_{in}} \sum_{m,n}^{N_{in}(x,y)} G_{in}(m,n) - \frac{1}{n_{out}} \sum_{m,n}^{N_{out}(x,y)} G_{out}(m,n),$$

where

$$G_{in}(m,n) = |f(m,n) - f(m+1,n)| + |f(m,n) - f(m,n+1)|,$$

and $G_{out}(m,n)$ is defined for the $N_{out}(x,y)$ neighborhood in the same manner. Finally, the $n_{in}$, $n_{out}$, $N_{in}$, and $N_{out}$ are identical to those defined for the previous feature.

The local variation feature, $F_{x,y}^3$, was chosen because human targets often show greater variation in temperature than natural objects in the background. This feature merely determines the average absolute difference between each pixel and the mean of the internal region (inner box), and compares it to the same measurement for a local background region (outer box). This feature at pixel $(x,y)$ is computed as:

$$F_{x,y}^3 = \frac{L_{in}(x,y)}{n_{in}} - \frac{L_{out}(x,y)}{n_{out}},$$

where

$$L_{in}(x,y) = \sum_{m,n}^{N_{in}(x,y)} |f(m,n) - \mu_{in}(x,y)|,$$

$$\mu_{in}(x,y) = \frac{1}{n_{in}} \sum_{m,n}^{N_{in}(x,y)} f(m,n),$$

while $L_{out}(x,y)$ and $\mu_{out}(x,y)$ are defined similarly for the $N_{out}(x,y)$ neighborhood.

For each input image, the strength of each feature is first computed for each pixel location. Normally, the dynamic range of these computed strengths varies between different features and different input frames. Therefore, these feature values are normalized across the image, feature by feature, so that the feature value at each pixel represents the number of standard deviations that the pixel stands apart from the values for the same feature across the image. With this normalization, the resulting confidence images from different input images are always occupying the same dynamic range; hence, a global acceptance threshold can be meaningfully applied across the whole data set. The normalized $k$-th feature value at pixel $(x,y)$ is calculated as:

$$F_{x,y}^k = \frac{F_{x,y}^k - \mu_k}{\sigma_k},$$

where

$$\mu_k = \frac{1}{S} \sum_{m,n}^S F_{m,n}^k, \quad \sigma_k = \frac{1}{S} \sum_{m,n}^S (F_{m,n}^k - \mu_k)^2,$$

and $S$ is the size of the feature image.

After normalization, the features images are linearly combined into a confidence image.
To determine the detection locations from the confidence image $\Phi$, the pixel location associated with the maximum confidence value is first chosen and designated as the first detection. Then the confidence values of all pixels within a human-sized neighborhood around the peak pixel are set to zero, so that the search for subsequent detections will not choose a pixel location corresponding to the same target again. The process is then repeated for a predefined number of detections or other stopping criteria. Figure 2 shows the Detalex outputs of six typical input images, in which each white box represents the location and size of a potential human target.

2.2. Second Stage: EIGMLP Clutter Rejecter

In the second stage of Figure 1(b), the EIGMLP uses an eigenspace transformation to perform the dimensionality reduction and a back-end MLP (BMLP) to reject clutter and accept targets. Figure 3 shows the structure of EIGMLP, which is a special feed-forward neural network constructed by concatenating an eigenspace transformation layer with a BMLP. The advantage of this concatenated structure is that it enables joint optimization between the transformation layer and the BMLP classifier, which is achieved by adjusting the corresponding weights of the transformation layer and the BMLP classifier, with a much smaller dimension (usually $k \ll n$) and a much larger separability between the classes, can then be sent to BMLP.

The inputs to EIGMLP are image patches (often referred to as image chips in literatures) extracted from input images based on Detalex results, which are depicted by the white boxes in Figure 2. The dimensionality of an input image chip is greatly reduced by projecting the chip onto a small set of eigenvectors and retaining only the projection values. These projection values are then non-linearly combined by BMLP to produce a new assessment on the likelihood that the given input chip might actually belong to a legitimate target.

Instead of using the unsupervised generalized Hebbian algorithm [4] to form the eigenspace transformation layer, the supervised two-class eigenspace separation transform (EST) is used. EST was shown to possess remarkable compaction and discriminative power in previous work for automatic target recognition [5]. The transformation matrix $S$ of the EST can be obtained as follows:

1. Compute the $n \times n$ correlation difference matrix

$$
\tilde{M} = \frac{1}{N_1} \sum_{p=1}^{N_1} x_{1p} x_{1p}^T - \frac{1}{N_2} \sum_{q=1}^{N_2} x_{2q} x_{2q}^T,
$$

where $N_1$ and $x_{1p}$ are the number of patterns and the $p$-th training pattern of Class 1, respectively. $N_2$ and $x_{2q}$ are similarly related to Class 2 (which is the complement of Class 1, such as clutter versus target).

2. Calculate the eigenvalues of $\tilde{M}$, $\{\lambda_i \mid i = 1, 2, \ldots, n\}$.

3. Calculate the sum of the positive eigenvalues, which is $E_\pi = \sum_{i=1}^{n} \lambda_i ; \forall \lambda_i > 0$; as well as the sum of the absolute values of the negative eigenvalues, which is $E_\nu = \sum_{i=1}^{n} |\lambda_i| ; \forall |\lambda_i| < 0$.

(a) If $E_\pi > E_\nu$, then take the $k$ eigenvectors of $\tilde{M}$ with positive eigenvalues to form the $n \times k$ matrix $S$.

(b) If $E_\pi < E_\nu$, then take the $k$ eigenvectors of $\tilde{M}$ with negative eigenvalues to form the $n \times k$ matrix $S$.

(c) If $E_\pi = E_\nu$, then use either subset of eigenvectors to form the matrix $S$, preferably the smaller subset.

Given the $S$ transformation matrix, the projection $y_p$ of an input pattern $x_p$ is computed as $y_p = S^T x_p$. The $y_p$, with a much smaller dimension (usually $k \ll n$) and a much larger separability between the classes, can then be sent to BMLP.

2.3. Third Stage: Evidence Integrator

As shown in Figure 1(c), the third and final component of the proposed algorithm is an evidence-combining MLP (ECMLP), whose function is to produce the final detection output by non-linearly combining the internal features of Detalex and the output of EIGMLP. The ECMLP is just another simple MLP, whose structure is very similar to that of the BMLP.
3. EXPERIMENTAL RESULTS

To examine the performance of the proposed multi-stage stationary human detection algorithm, a challenging set of 1,763 FLIR still images was used. Each image measured 324 x 256 pixels at 14-bit resolution, with one to four human targets posed at distances ranging from 2–80 m. Due to the huge variation in range, the size of human targets varies greatly across the images, ranging from a few pixels to covering a third of the image. The difficulty is also increased by the diverse scenes, cluttered backgrounds, varying poses and very low contrast of the human targets, as well as major occlusions by vegetation, cars, and buildings. The 1,099 training images and 664 testing images were chosen based on separate dates of data collection. Figure 2 shows six of the typical images in the training set.

First, the Detalex was used to perform the detection task, which is the configuration depicted in Figure 1(a). The Detalex operational feature set included the maximum gray level, contrast-box, gradient strength, and local variation features. These features at each pixel location were first individually computed and normalized, and then linearly combined to produce the confidence value at that point. The detection results were compared to the ground-truth location and size of the corresponding targets. By varying the acceptance threshold based on confidence values, thus changing the hit rate and false alarm rate, the receiver operating characteristic (ROC) curves were plotted for the training and test sets. As shown in Figure 4, the horizontal axis of these plots represents the average number of false alarms per frame, while the vertical axis represents the percentage of all legitimate targets in the data set that were correctly detected at various acceptance thresholds. As shown by the lowest pair of ROC curves in Figure 4, Detalex gradually achieved the maximum hit rates of 97.13% and 97.53% at 8.71 and 8.78 false alarms per frame for the training and testing set, respectively. At 1 false alarm per frame rate, however, Detalex correctly located only 28.41% and 30.46% of the training and testing targets, respectively.

The performance of EIGMLP clutter rejeter was then examined using the image chips identified by Detalex, which is the setting represented by Figure 1(b). Detalex-detected image chips were scaled to a common size of 800 pixels, and the top 15 EST eigenvectors were used for the transformation layer of EIGMLP. As shown by the middle pair of ROC curves in Figure 4, EIGMLP significantly improved the detection performance at the operating zone that permits very few false alarms. At 1 false alarm per frame rate, however, Detalex correctly located only 28.41% and 30.46% of the training and testing targets, respectively.

Finally, the performance of the complete algorithm was examined, as depicted in Figure 1(c), by non-linearly combining the Detalex features and the output of EIGMLP to form the final detection decision. The top pair of ROC curves in Figure 4 clearly shows that ECMLP achieved remarkable performance improvements over the two previous architectures at any false alarm rate. The ECMLP training and testing hit rates at 1 false alarm per frame rate were 80.62% and 72.02%, respectively.

4. CONCLUSIONS

A multi-stage stationary human detection algorithm is proposed and evaluated for FLIR imagery, in which the computational resources are gradually allocated to find the potential human targets in still images. The results show a significant improvement in detection performance at any false-alarm rate by cascading the Detalex, EIGMLP, and ECMLP together.

5. REFERENCES


