ROBUST TARGET DETECTION AND TRACKING IN OUTDOOR INFRARED VIDEO

Changchun Li, Nan Jiang, Jennie Si

Department of Electrical Engineering
Arizona State University

Glen P. Abousleman

General Dynamics C4 Systems
Scottsdale, AZ 85257

ABSTRACT

Automated tracking of targets within outdoor infrared (IR) video sequences poses a host of challenges. These include automatic gain adjustment in the IR camera, extreme granularity, large luminance changes, and uncontrolled environmental factors such as moving foliage, animals, and birds, among others. To address these problems, we present an IR video target tracking system for stationary cameras that learns and divides the video frames into reliable and unreliable regions. A difference-frame-based method can recognize moving regions with high sensitivity and reliably discern background clutter from target motion. A low-complexity target validation process is presented, which in conjunction with the reliable region masking, dramatically reduces the number of false alarms. We demonstrate the outstanding performance of the proposed system using real-world IR video sequences with difficult background motion clutter, as well as with small and blurred moving targets.

Index Terms— target detection, target tracking, infrared video, reliable region, tracking-based detection

1. INTRODUCTION

Detection and tracking of targets in infrared (IR) video have been the subject of research for many years. Infrared video often has low contrast, strong thermal noise, and poor resolution. Furthermore, in outdoor environments, the uncontrolled lighting, strong foliage movement, flowing water, and adverse weather are severe challenges faced by detection and tracking algorithms. Depending on the target size, infrared tracking can be classified as point-based and region-based. Point-based approach often model targets as a two-dimensional Gaussian function, e.g.,[1], [2]. Longin et al [3] proposed a detection algorithm by examining the changes in spatiotemporal texture. They used nearest-neighbor methods for tracking that resulted in significantly better performance over the Stauffer-Grimson approach. Mei el al [4] presented an algorithm to classify vehicles in IR video by the integration of detection, tracking, and recognition. A simple detection method was used prior to particle filtering for tracking and classification. However, their detection method is vulnerable to environmental disturbances such as background clutter. The wavelet transform is used to suppress the background in [5], and the threshold is chosen to maximize inter-class variance to achieve better detection, which is suitable for long-range small IR target detection. Zhang et al [6] used morphological operators to estimate the background and introduce target energy features to detect IR targets, which can be applied to long-range targets with a constant background, e.g., sea or sky, but is not useful for IR video with close targets.

Since no solution for detection and tracking exists for close-view outdoor IR video, where environmental disturbances are significant, we propose a new tracking-based detection approach for the detection and tracking of dim IR targets. The core algorithm divides the input frame into reliable and unreliable regions automatically, which increases its sensitivity to detecting small and dim IR targets, and greatly reduces false alarms. A tracking-based detection scheme is used to recognize true moving targets and exclude those falsely detected blobs caused by noise, luminance changes, or heavy background clutter. A modified nearest-neighbor data association method is used for tracking. Excellent performance with a very low false alarm rate, high sensitivity of detection, and consistent tracking are demonstrated with real-world outdoor IR video sequences.

The remainder of this paper is organized as follows: Section II presents the background model; Section III provides a detailed description of the tracking-based detection and tracking algorithm; experimental results are presented in Section IV; and a conclusion is given in Section V.

2. BACKGROUND MODEL

Since we assume a static IR camera, it is natural to use frame differencing for detecting moving targets. However, frame differencing between consecutive frames itself cannot provide satisfactory results when the target of interest does not show significant movement in the direction parallel with the image plane of the camera. If a background image without any moving targets is available, one can get a better result by computing the difference with the current frame. Rather than use either consecutive frame differencing or background differencing only, we use both methods concurrently because

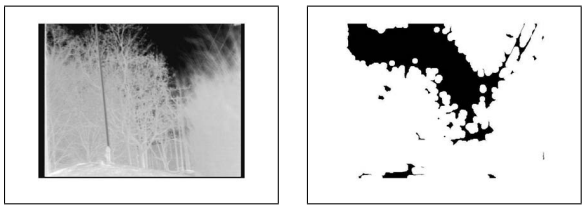
The research was supported by General Dynamics C4 Systems.

they provide complementary results. The consecutive frame differencing approach can capture the real-time change between frames, while background differencing gives the necessary complement when a moving target stops moving or moves very slowly during the detection and tracking.

In outdoor environments, many false alarms occur in areas with swaying trees, flowing water, and sharp edges. If we can identify these areas and mark them as unreliable regions, we can simultaneously improve the detection accuracy and lower the false alarm rate. First, a rough training procedure is performed to obtain a smoothed background and pixel variance. Secondly, a refining process is applied over the difference image to find more unreliable regions. Assuming there are no moving targets during the training interval, any deviations will result from environmental factors. A moving window represented by equations (1) and (2) is used to update the background and pixel variance adaptively with time:

$$B_t = (1-\alpha)B_{t-1} + \alpha I_t(i,j) \quad (1)$$

$$\delta_t = sqrt((1-\beta)\delta_t^2 + \beta(I_t(i,j) - \mu_t)^2). \quad (2)$$



(a) Trees: original frame (b) Trees: after edge and close

Fig. 1. Example: "trees" pattern.

The above procedure is comparable to true time averaging and is more suitable for real-time applications. Following the rough training process, areas with a large pixel variance are marked as unreliable regions. In the refining process, a difference image is obtained by a combination of consecutive frame differencing and frame differencing with the trained background. Morphologic edge and close are taken over the entire difference image to produce a binary image. When there are environmental disturbances, large errors are generated in the difference image. Accordingly, some patterns can be clearly identified on the binary image. These areas will be marked as unreliable regions as well. An example of a "trees" pattern is shown in Figure 1. When converting the difference image into a binary image, one can choose direct threshold or edge operators. The benefit of the edge operator is that it can capture structural information. For example, the close operator connects isolated parts into continuous regions. Thus, the unreliable region's structure is readily identified. A mask is generated to represent the unreliable regions identified from both the rough training and refining processes. The aforementioned process is described by Algorithm 1.

Algorithm 1 Unreliable region detection

Stage 1: Rough training
with frame I_t, where $t = 1 \cdots t_0$
Obtain $B(i,j)$ and $\delta(i,j)$ by equations (1) and (2)
Stage 2: Refining
$\mu_h = \frac{1}{M*N}\sum_{i,j}\delta(i,j)$, $\delta_h = sqrt(\frac{1}{M*N}\sum_{i,j}(\delta(i,j)-\mu_h)^2)$,
$Mask = \mathbf{0}_{M\times N}$, M and N are image height and width respectively,
If $|\delta(i,j) - \mu_h| \geq \gamma * \delta_h$, $Mask(i,j) = 1$,
where γ is a constant, 3 is taken here,
for frame $I_t(i,j)$, $t = t_0 + 1 \cdots t_1$,
$D_t(i,j) = (I_t(i,j) - I_{t-2}(i,j) + I_t(i,j) - B(i,j))/2$,
Get edge image $Edge(i,j)$ from $D_t(i,j)$ by "canny" method,
with thresholds 0.1 and 0.25,
Morphologic close, structuring element 'disk' of radius r,
If $\sum_k S(k) > \eta * M * N$, $t = t+1$ go to next frame,
where $S(k)$ is a connected region, η is a constant, 0.8 is taken here,
otherwise,$Mask = Mask \cup S(1) \cup \cdots \cup S(K)$,
Merge all K regions into $Mask$.
end

3. TRACKING-BASED DETECTION AND TRACKING

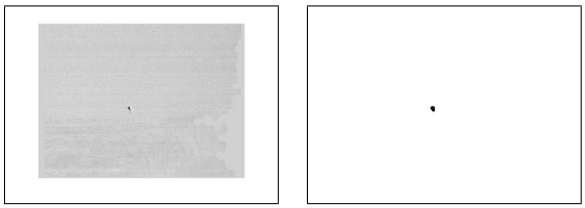
As stated previously, the sensor noise and large background clutter in uncontrolled outdoor environments (such as trees, grasses, or rivers) are the main sources of false alarms if the binary image is generated directly from the difference image. Even with a good background model, one cannot ensure that all false alarms are suppressed. Accordingly, we propose a tracking-based detection and tracking scheme to suppress the false alarms. Its main idea can be summarized as follows: Prior to being claimed as a new target, a candidate is put into a buffered queue until its motion pattern complies with predefined criteria. After a difference image is extracted from each input video frame, a reliable-region mask is applied to remove the unreliable regions. Rather than use the raw difference image, a gradient image is generated. Because the gradient is the derivative of the raw difference image, it is invariant to global luminance change and is more sensitive than the raw difference image in capturing the dim targets in IR video. To remove the artificial edges that remain following the removal of the unreliable regions, a shrink mask generated from the reliable-region mask is overlaid on the gradient image. As shown in Figure 2, the unreliable region is erased and the dim small target is identified clearly. The detailed procedure is presented in Algorithm 2.

Once the candidates are detected with the above process, the tracking-based detection and tracking are employed as follows: First, newly detected candidates in the current frame try to set up associations with targets in a true target pool. If no association is found, they try to set up associations with targets in a potential target pool. If still no association exists, the candidates are put into a potential target pool as new potential targets. A vitality counter is assigned to each candidate in the potential target pool. If a candidate in the potential target pool is updated, its corresponding vitality counter is increased as well; otherwise, its vitality counter decreases. If an object's

Algorithm 2 Detect candidates

With background model $B(i,j)$ and reliable-region mask $Mask(i,j)$ ready,
for frame $I_t(i,j), t = t_1 + 1$ to T,
$L = \mathbf{0}_{M \times N}$,
$D_t(i,j) = (I_t(i,j) - I_{t-2}(i,j) + I_t(i,j) - B(i,j))/2$,
Morphologic edge: $Edge(i,j)$ from $D_t(i,j)$ by "canny" method,
with low and high thresholds 0.1 and 0.25 respectively,
Morphologic:close, structuring element 'disk' of radius r,
If $\sum_k S(k) > \eta * M * N$, $t = t + 1$ go to next frame,
where $S(k)$ is the area of connected regions,
η is a constant, 0.8 is taken here,
Apply reliable-region mask, $\bar{D}_t = D_t * Mask$,
Get gradient of $\bar{D} \Rightarrow G_t$,
Get the shrink mask $\bar{V}$ from $Mask$, then $\bar{G}_t = G_t * \bar{V}$,
$m_G = (1/M * N) \sum \bar{G}_t$,
$\sigma_G = sqrt(1/(M * N - 1) \sum (\bar{G}_t - m_G)^2)$,
If $|\bar{G}_t - m_g| \geq \lambda * \sigma_G$, $L(i,j) = 1$.
Detected candidates=connected regions of $L(i,j)$,
end

vitality counter decreases to zero, it will be removed from the potential target pool. If a target's vitality counter reaches an upper threshold, its identity is checked to decide whether or not it is a true moving object. If a candidate passes the testing, it will be put into the true target pool. For the targets in the true target pool, they will be compared with the newly detected targets to decide if an association exists. If an association exists, the target's vitality counter is increased; otherwise, the counter decreases. When a true target's vitality counter reaches zero, it will be removed from the true target pool. If a candidate is a true moving object, it is able to maintain its appearance and moving pattern consistently for at least some period of time. The temporary occurrence of a false alarm won't last long enough to be classified as a true target. Thus, a comprehensive detection and tracking system is formed, which can deal with newly birthed object, tracking, and disappearing objects in a consistent way.



(a) Difference image of reliable region

(b) Detected target using gradient image

Fig. 2. Detect possible candidates.

In the process of data association, due to the large noise content in IR video, there are often multiple candidates in the neighboring region of a true object. Unlike video collected by EO sensors, there is no stable target appearance model in IR video. Additionally, the objects are often occluded by natural objects such as trees, buildings, and rocks. All of these factors make it difficult to extract stable and reliable features. Accordingly, we propose a multiple feature vector weighted by corresponding long-term statistics. In our system, we use six features including velocity direction and magnitude, bounding box width and height, mean of object intensity, and texture. We use the following method to weight different features automatically. The distance measure between a target and a candidate is computed using Equation (3):

$$S = (1/W) \sum_i w_i * exp(-0.5 * (\Delta x_i)^2 / \sigma_i^2), \qquad (3)$$

where $w_i = \frac{\sigma_{0i}^2}{(\Delta x_i)^2 + \sigma_i^2}$ is a weighting factor, $W = \sum_i w_i$ is a normalization factor, σ_{0i} is the predefined standard deviation, σ_i is the current standard deviation, and Δx_i is the relative difference for the i^{th} feature: $\Delta x_i = \frac{x_{t-1} - x_t}{max(x_{t-1} - x_t)}$. If the computed score, S, is greater than a threshold, S_0, but less than a higher threshold, S_1, the candidate is associated with the object, but the appearance model, bounding box, mean, and texture are not updated. If the score is higher than S_1, the appearance model is updated, and each feature's current variance is updated as well using equation (4):

$$\sigma_i = sqrt((1 - \beta) * \sigma_i^2 + \beta * (\Delta x_i)^2). \qquad (4)$$

The above weighting method benefits the reliable features while suppressing the unreliable ones. The output can be regarded as a probability and can be used within any probability-based framework. Since all the features should have zero mean according to our design and assumptions, if a feature is stable, σ_{0i} will be close to σ_i, and Δx_i will be close to zero. That feature's weighting, w_i, will be close to one. Similarly, if one feature becomes unstable, its current variance will increase significantly with time. Thus, both $(\Delta x_i)^2$ and σ_i become large, and the weighting factor, w_i, will be close to zero. That feature's importance will be reduced greatly from the total measurement.

To identify whether a target in the potential target pool is a true moving target, a motion pattern analysis is performed. False alarms tend to have positions that are randomly distributed, while true moving objects are statistically stable along some path. A fast pattern analysis computes the average angle change of the moving target and the average moving distance. If the average turning angle is smaller than a threshold, θ_0, and the moving distance is greater than a threshold, d_0, then we perform a 3^{rd}-order polynomial curve fitting and find the correlation coefficient between the observed position and the estimated position. A 1^{st}-order autocorrelation coefficient is also computed. If both correlation coefficients are greater than a threshold, c_0, it is recognized as a true object and put into the true target pool for continued tracking. An example of path fitting for a true moving target is shown in Figure 3.

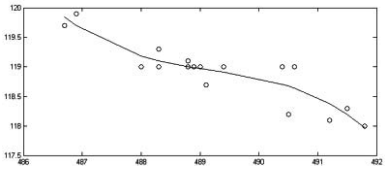
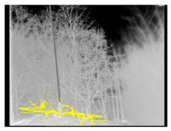


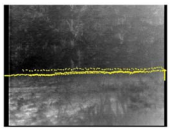
Fig. 3. Motion pattern example.

4. EXPERIMENTAL RESULTS

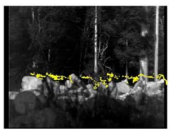
In this section, we present the implementation and test results of the proposed system. The video sequences were generated by an unattended IR camera capturing various outdoor scenes. All sequences have a frame size of 640×480 pixels and a frame rate of 10 frames per second (fps). The system performs very well on all eleven video sequences. Due to space limitations, we present the results for only three sequences.



(a) Tracking result with "trees" sequence



(b) Tracking result in "mt04" sequence



(c) Tracking result in "mt05" sequence

Fig. 4. Tracking examples.

A median filter was applied to reduce the noise, and 30 seconds of video was used for rough training and refining. The tracking paths (yellow) are overlaid over the background image to show the results clearly. Figure 4(a) is taken in dense woods, where the foliage and tall grass occupy a large portion of the scene. A person enters the scene from the corner and crosses the woods through a curved path. The person is often partially or fully occluded by the woods or large surface bumps. Although regions close to the black sky and tall grass cannot be identified as unreliable regions, and quite a few blobs are detected as potential candidates, our algorithm can differentiate them from the true moving target without any false alarms. Figure 4(b) is taken by a camera hidden in the tall grass. The grass sways in the wind and creates pronounced disturbances. The tracking subject, a motorcyclist, is detected quickly and tracked accurately for the entire sequence duration with no false alarms. Figure 4(c) is an IR video sequence taken in front of rocky area during the daylight. The strong and gusty wind enhances the disturbances from the swaying trees and grass. The tracking subject is a person who crosses the rocky area. He has frequent changes in his motion pattern and shape when standing, walking, climbing, stooping, and jumping. The system accurately tracks the subject until he crosses the entire rocky area and disappears from view.

5. CONCLUSIONS

We have presented a novel video tracking system for outdoor infrared video applications. Region masking and buffered target detection and tracking processes were introduced to dramatically reduce the number of false alarms with complex background clutter. Performance results were presented with real-world outdoor IR video sequences. It was shown that excellent target detection and tracking could be obtained while simultaneously rejecting the false alarms due to extreme background motion clutter.

6. REFERENCES

[1] K.L Anderson and R.A Iltis, "A tracking algorithm for infrared images based on reduced sufficient statistics," *Aerospace and Electronic Systems, IEEE Transactions on*, 1997.

[2] Y. Xiong, J. Peng, M. Ding, and D. Xue, "An extended track-before-detect algorithm for infrared target detection," *Aerospace and Electronic Systems, IEEE Transactions on*, 1997.

[3] L.J Latecki, R. Miezianko, and D. Pokrajac, "Tracking motion objects in infrared videos," in *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005, pp. 99–104.

[4] X. Mei, S.K Zhou, and H. Wu, "Integrated detection, tracking and recognition for ir video-based vehicle classification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. V–V.

[5] Y.Q Sun, J.W Tian, and J. Liu, "Background suppression based-on wavelet transformation to detect infrared target," in *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, 2005, pp. 4611–4615.

[6] S.J Zhang, Z.L Jing, J.X Li, and H. Leung, "Small target detection of infrared image based on energy features," in *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, 2003, pp. 672–676.