A Real-time, Automatic Target Detection and Tracking Method for Variable Number of Targets in Airborne Imagery

Tunç Alkanat, Emre Tunali and Sinan Öz
Image Processing Department, ASELSAN Inc. Microelectronics, Guidance and Electro-Optics Division, Ankara, Turkey

Keywords: Real-time Target Detection, Multiple Target Tracking, Temporal Consistency, Data Association, Target Probability Density Estimation, Adaptive Target Selection.

Abstract: In this study, a real-time fully automatic detection and tracking method is introduced which is capable of handling variable number of targets. The procedure starts with multiple scale target hypothesis generation in which the distinctive targets are revealed. To measure distinctiveness; first, the interested blobs are detected based on Canny edge detection with adaptive thresholding which is achieved by a feedback loop considering the number of target hypotheses of the previous frame. Then, the irrelevant blobs are eliminated by two metrics, namely effective saliency and compactness. To handle the missing and noisy observations, temporal consistency of each target hypothesis is evaluated and the outlier observations are eliminated. To merge data from multiple scales, a target likelihood map is generated by using kernel density estimation in which weights of the observations are determined by temporal consistency and scale factor. Finally, significant targets are selected by an adaptive thresholding scheme; then the tracking is achieved by minimizing spatial distance between the selected targets in consecutive frames.

1 INTRODUCTION

Multiple target detection and tracking has significant importance for many applications, including reconnaissance and surveillance in which the major goal is to reveal trajectories of the targets throughout the scenario. Considering the recent developments, many electro-optical systems are in need of full automation for achieving this task. Therefore, many multi-tracking algorithms include two fundamental stages as the automatic, time independent detection of targets; and association of the detections in the temporal space. Although there exists many research on the subject (Berclaz et al., 2011; Niedfeldt and Beard, 2014; Andriyenko and Schindler, 2011), problem remains to be challenging mainly due to unknown and changing number of targets; noisy and missing observations; interaction of multiple targets. Moreover, all these challenges are needed to be solved in a time efficient manner for real-time applications.

The outstanding target detection concept can be interpreted in different ways and many interest point detection techniques can be used as a starting point to determine such objects on an image. In the literature, there exists numerous interest point detection methodologies based on blob detection (Lowe, 2004; Bay et al., 2008), corner detection (Harris and Stephens, 1988; Rosten and Drummond, 2006) and edge detection (Canny, 1986; Prewitt, 1970; Sobel and Feldman, 1968). Rather than searching for corners or blobs, defining the outstanding object from the contrast is a better choice for our application since we are not only interested in cornered or blob-like structured objects. In this sense, usage of edge detectors yields better generalization and among edge detections methods Canny edge detection shows its superiority due to its ability of generating closed contours by merging weak edges with the strong edges around their vicinity. Furthermore, the low computational cost of the Canny edge detector also allows the usage of pyramid structure in order to respond targets in different scales without introducing any restriction for real-time processing which is one of the major goals.

Another important aspect of the detection phase is determining the number of targets dynamically since the selection of predetermined number of targets would be problematic. To be clearer, if the number of targets is smaller than the expected number of targets, the system is forced to introduce insignificant targets to the track list. Likewise, in the scenes having higher number of significant targets than the expected, some of the significant targets will be ignored. To deal with

Alkanat T., Tunali E. and Öz S..
A Real-time, Automatic Target Detection and Tracking Method for Variable Number of Targets in Airborne Imagery.
DOI: 10.5220/0005298400610069
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 61-69
ISBN: 978-989-758-090-1

Figure 1: Sample outputs of the proposed solution demonstrating successful tracking for variable number of targets.

the unknown, changing number of targets and develop an unsupervised approach, a target selection procedure is also introduced.

Temporal association of detections is the fundamental problem of multi-target tracking. Despite existence of many detection methodologies, none of the detection methods can provide robust detection results to be used in the data association stage. To be more precise, detections may be misleading from time to time and the outlier data should be handled while achieving the data association. For this purpose, one of the most popular and well studied method is Kalman filtering (Kalman, 1960) which deals with the outliers by achieving a compromise between the probabilistic model of the target motion and the measurement. Although this methodology is effectively used in many applications (Tsai et al., 2010; Ramakoti et al., 2009), requirement of the predetermined motion model becomes a significant restriction. Usage of particle filters (Ristic et al., 2004) can address some of the limitations of the Kalman filters by exploring multiple hypotheses; however this results in an increased computational complexity. Other widely used techniques for the association problem are joint probability density association filters (JPDAF) (Fortmann et al., 1980) and multiple hypothesis tracking (MHT) (Reid, 1979). The JPDAF actually uses soft data assignment by considering the probability of a measurement belonging to more than one track which results in a single hypothesis for summarizing all the previous measurements. The main limitation of JPDAF is the assumption on number of targets which is stated to be fixed. Hence, it is not capable of handling targets entering/leaving the scene. In MHT, this problem is solved by integrated track initiation. Association algorithm of MHT is a hypothesis based brute force implementation which aims to generate all possible hypotheses and requires high computational load. Moreover, MHT also requires a large memory space to be used; since the different hypotheses from previous frames are kept in the memory. Instead, the proposed method obtains measurements with a pyramid structure and benefits from motion heuristics together with a probability density estimation methodology which is designed for merging measurements from different levels of the observation pyramid. The density estimation method is based on Parzen windowing (Parzen, 1962), and benefits from a weighting scheme to tolerate missing and noisy observations with low computational cost.

The rest of the paper is organized as follows: The proposed target detection and tracking method is explained in Section 2, the conducted experiments are analyzed in Section 3, and finally the study is concluded in Section 4 where discussions are made.

2 PROPOSED METHOD

The multiple target detection and tracking method proposed in this paper consists of 4 main steps: First, target hypotheses are generated for different scales based on distinctiveness and compactness assumptions of target model, then temporal consistency of each target hypothesis is calculated for both rejecting outliers and compensating missing detections in a time efficient manner. By using these consistent target

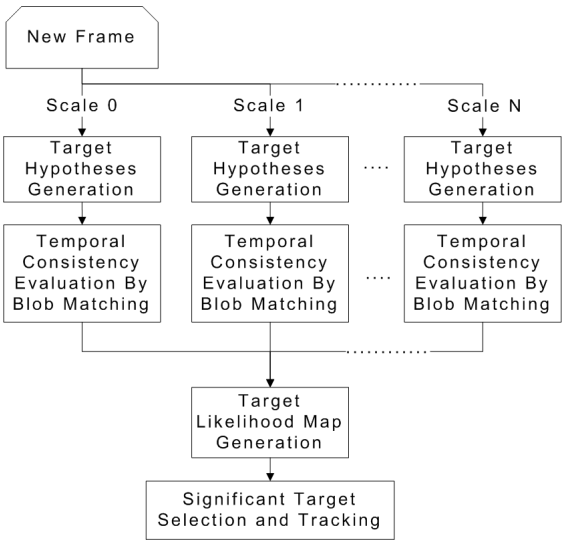

Figure 2: General overview of the proposed solution.

hypotheses, from each scale of the observation pyramid, a target likelihood map is generated representing the target existence likelihood at each pixel. Finally, outstanding (relevant) targets are selected form the likelihood map by using an adaptive thresholding scheme and selected targets are associated in consecutive frames to reveal their trajectories.

2.1 Target Hypotheses Generation

To achieve automatic target detection, each target candidate fulfilling some preliminary requirements should be further analyzed to decide whether it is a relevant target or not. The target candidates are referred as target hypotheses and generated at each scale of the observation pyramid, obtained by downsampling the original frame, separately. Therefore, for both hypotheses generation and selection, some assumptions are made to describe the target model.

The first assumption is the distinctiveness assumption stating that target candidates should be distinctive from their surroundings. Actually, this assumption is made based on human visual attentional system in which robust saliency detection mechanisms provide focus of attention to the salient regions pre-attentively for further processing. Again similar to human visual system, the distinctiveness is measured by the intensity difference. Most of the saliency detection methods are founded on the same principle; however saliency detection in global scale (by considering the whole scene) would generally require high processing time which may not be suitable for real-time applications. Since the computational complexity is one of the key issues, target hypothesis generation procedure starts with edge detection which is a simple way of detecting contrast between neighboring pixels. For edge detection, Canny edge detector is preferred for both its low computational complexity and capability of generating closed contours by merging weak edges with the strong edges around their vicinity. After employing Canny edge detection, morphological closure (to increase probability of generation of closed contours) and filling operations are performed on edge map to obtain the possible target blobs. The importance of filling operation becomes more prominent when a possible target has a layered structure, having nested closed contours inside the target as in Fig. 3 in which an inner loop is detected due to the reflection of the daylight. In such a scenario, detection of the complete vehicle is more preferable than detection of the spot as a separate target; and filling the closed contours inherently yields the selection of the outer most closed contour since both the inside of the spot and the vehicle are filled. After morphological operations, centroids of the resulting filled blobs are obtained by using connected component analysis.

Figure 3: Effect of filling on a target with layered contours

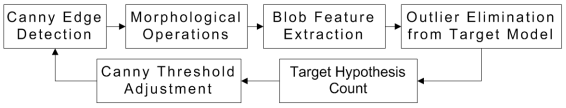

Figure 4: Flowchart for target hypotheses generation

Usage of static thresholding in Canny edge detection can be problematic since different scenes may have different contrast spans. Therefore, while a static threshold can satisfactorily detect targets in scenes having high contrast, it may fail to disclose any edges in scenes having low contrast in which targets are still visible to the human visual system. Since the aim is to detect relatively high intensity differences, a dynamic thresholding scheme is applied in which Canny thresholds are adjusted dynamically with a feedback loop, Fig. 4, whose input is the target hypothesis count from the previous frame. To achieve dynamic thresholding, high threshold of the Canny is simply decreased/increased with a certain amount if the target hypothesis count is less/higher than the desired number of hypothesis. In this manner, dynamic thresholding provides another advantage which is keeping the number of blobs and thus targets within a limit.

Although edge detection reveals regions with relatively high contrast from its surroundings, it can only give some insight about the distinctiveness level of the target. To mathematically represent distinctiveness of the target candidates, a new metric referred as effective saliency is introduced based on a saliency detection methodology (Wei et al., 2012) in which the saliency problem is tackled from a different perspective by focusing on background more than the object. Although there are various saliency detection algorithms (Hou and Zhang, 2007; Achanta et al., 2009; Cheng et al., 2011), the main motivation of using this method is its capability of extracting a saliency map with low computational cost. However, usage of this technique is restricted with the boundary assumption which is the reflection of a basic tendency that a cameraman does not crop salient objects in the frame. Thus, the image boundary is assumed to be background. Satisfying the assumption, the salient regions are determined by identifying the patches with high geodesic distance to the image boundaries. For the calculation of geodesic distance, definitions of (Wei et al., 2012) is followed and the image is divided into vertices which are composed of inner

patches P_i and background nodes (B, image boundaries). Hence, two types of edges: internal edges, connecting all adjacent patches; and boundary edges, connecting image boundary edges to the background node are obtained ($\xi = (P_i, P_j | P_i \textit{ is adjacent to } P_j) \cup (P_i, B | P_i \textit{ is on the image boundary})$). Then, the geodesic saliency of a patch P is calculated by accumulating edge weights (intensity differences) along the shortest path from P to virtual background node B in an undirected weighted graph as given in Eqn.1,

$$Saliency(P) = \min_{P_1=P,P_2,\ldots,P_n=B} \sum_{i=1}^{n-1} weight(P_i, P_{i+1}). \quad s.t.(P_i, P_{i+1}) \in \xi \tag{1}$$

In our case, the boundary assumption of (Wei et al., 2012) is fulfilled by calculating saliency map from the image patches that are co-centered the blobs obtained from Canny edge detection and that encapsulate objects with their immediate surrounding. Actually, selection of the image patch is the first step of calculating effective saliency metric. After calculating the saliency map, a binarization threshold is obtained by using Otsu's thresholding (Otsu, 1979). Then, the effective saliency ($E_s(t)$), is calculated for each blob as in Eqn. 2 where dominant components (D.C.) represent the pixels whose saliency values are greater than the binarization threshold and $S_{blob}(x)$ is the saliency map obtained for each blob. High distinctiveness is a significant sign of a possible target, hence candidates that do not have a certain level of distinctiveness are eliminated.

$$E_s(t) = \frac{\sum_{x \in D.C.} S_{blob}(x)}{\sum_{y \in S_{blob}} S_{blob}(y)} . \tag{2}$$

Another assumption that is made for target model is the compactness assumption. Since the Canny edge detection reveals not only edges of the objects but also edges belonging to structural details in the scene, some of the detected blobs must be eliminated. Therefore, a further selection procedure is applied to the blobs satisfying the distinctiveness assumption. To achieve the task, the compactness metric is used which is actually nothing but a scalar specifying the proportion of number of the pixels belonging to blobs to the area of the minimum sized bounding box encapsulating the blob. Using this feature, the targets having degraded from rectangular shape are eliminated. This procedure is illustrated in Fig. 5. The remaining blobs satisfying both distinctiveness and compactness assumptions are referred as target hypotheses and further processed to find out their temporal consistency.

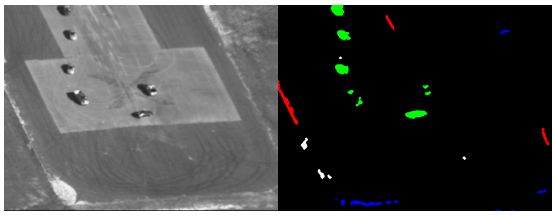

Figure 5: On left, original image. On right, blob mask of the original image with relevant targets, inconsistent targets, blobs violating compactness, blobs violating distinctiveness marked with green, white, red and blue, respectively.

2.2 Temporal Consistency Evaluation by Blob Matching

Although Canny edge detection is one of the simplest methods for contrast detection; it is vulnerable to noise and consequently becomes a source for noisy observations. More precisely, Canny edge detection may fail to provide closed contours, yielding missing observations in some frames or can produce artificial closed contours due to noise. Generation of faulty observations is a common problem and in some well-known techniques (Kalman, 1960; Fortmann et al., 1980), solution is based on probabilistic model on target behavior. However, this would be over-restricting for our problem since dealing with moving cameras generally result in complex motion patterns. Thus, for rejection of outliers and handling missing data, proposed method identifies an observation point as a target hypothesis if and only if the observation point keeps its presence for multiple frames. In other words, temporal consistency of a target is assured based on a scoring scheme in which higher score of a target hypothesis represents higher reliability.

The proposed scoring scheme is applied at each frame and starts with associating newly generated and existing target hypotheses. At first, for each new target hypothesis, existing hypotheses are searched in a neighborhood to satisfy the motion heuristic known as maximum velocity, (Yilmaz et al., 2006). Usage of such a simple model is both less restricting and requires much less computational load compared to other motion models. Existence of a match is decided by minimizing the norm of vectors that contain spatial distance and mean intensity difference of a new hypothesis to existing target hypotheses within the neighborhood. If match is found, the score of the matched existing target is increased. After matching all new target hypotheses, the score of the remaining (unmatched) existing target hypotheses are decreased. Then, unmatched new target hypotheses are considered as possible new targets entering the scene and initial scores are assigned according to their similarity

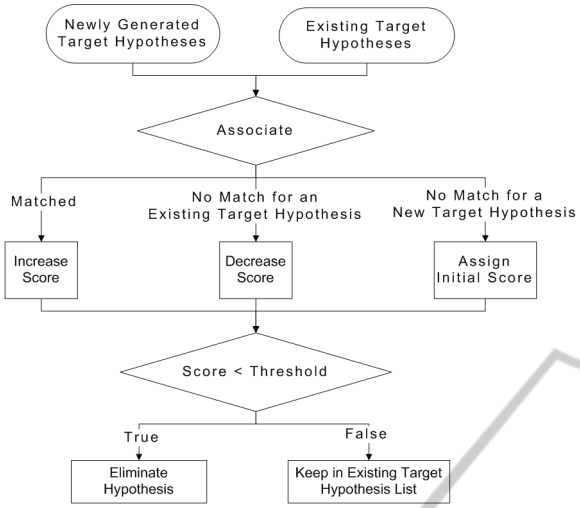

Figure 6: Proposed scoring scheme

to the target model description which is measured by the effective saliency metric. After adding new target hypotheses to the existing target list and adjusting the scores, target hypotheses list is reconstructed by eliminating the ones that below the score threshold. Following the scheme, the missing observations for limited number of frames would be tolerated since they are still considered as target hypotheses until their scores go below the threshold. In a similar fashion, the observations that are generated due to noise will also be eliminated within a limited number of frames since they are not persistent. On the contrary, new targets entering the scene will be considered as target hypotheses given that they are consistent. The proposed scoring scheme is summarized in Fig. 6.

2.3 Target Likelihood Map Generation

An important problem introduced by Canny edge detector is the false partitioning of a single object into multiple closed contours which is due to a failure in detecting the outermost contour of an object as a closed contour. This problem would result in multiple target initialization for a single object and appears more frequently for large sized objects due to the nature of the edge detector. Obviously, usage of the data provided by each scale of the pyramid together would definitely decrease the occurrence rate of the problem. Actually merging the data of different scales can be considered as a probability density estimation problem whose solution identifies the target likelihood map representing the existence probability of a target at each pixel.

Since no prior information exists about the target probability distribution, estimation is preferred to be achieved based on a non-parametric approach. To achieve this task, kernel density estimation (Parzen window method (Parzen, 1962)) is employed in which normal distribution is selected as the kernel function. Normal distribution is preferred assuming that effect of a target hypothesis on neighboring pixels yields a normal distribution whose peak is located on the centroid of the target hypothesis. In this manner, the variance of the normal distribution will determine the distance between the centroids of different target hypotheses to be merged.

To generate the target likelihood map, different from classical Parzen windowing, data is weighted with respect to its significance that is defined by two scalars which are temporal consistency and scale weights. Since the significance of a target increases with its temporal consistency, consistency weight (w_c) is obtained by the score whose calculation is explained in Sec. 2.2. Therefore, while decreasing the effect of mis-detected hypotheses from one scale of the pyramid, the weights of the corresponding target hypotheses are increased at the relevant scale yielding better localization. The second scalar, scale weight (w_s) is designed to select the importance of different scales of the pyramid. Since the partitioning occurs generally for the large objects; to compensate the erroneous data, detections obtained from lower resolutions (downsampled by a higher factor) are weighted proportional to the downsampling factor. The formal definition of the target likelihood for each pixel (x,y) is given in Eqn. 3,

$$P(x,y) = \frac{\sum_{j\in H} w_c \cdot w_s \cdot exp\left(-\frac{(x-x_j)^2+(y-y_j)^2}{2\sigma^2}\right)}{\sum_{\forall pixels}\sum_{j\in H} w_c \cdot w_s \cdot exp\left(-\frac{(x-x_j)^2+(y-y_j)^2}{2\sigma^2}\right)}, \quad (3)$$

where (x_j, y_j) is the locations from the set of target hypotheses H.

2.4 Target Selection and Tracking

Once the target likelihood map is obtained, target selection becomes nothing but a threshold selection problem which determines the lowest probability in the target likelihood map that will be considered as a target. Although the simplest solution is to use static threshold; dynamic thresholding is preferred due to the utilized scoring scheme applied to the target hypotheses. To achieve the task, the dynamic thresholding methodology proposed by (Aytekin et al., 2014) is followed which is designed to reveal distinctive intensity falls on a given image. This method analyzes the relationship between the local maxima of input image and the threshold is calculated using weighted

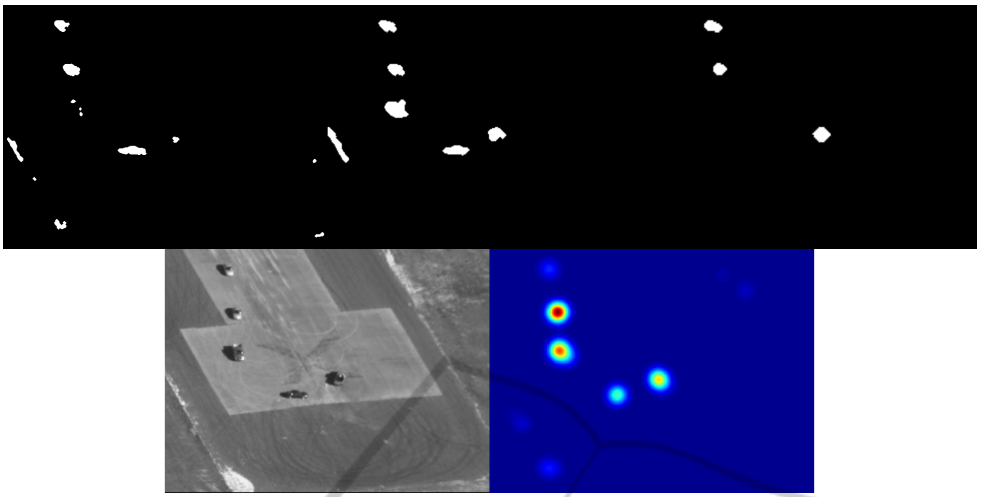

Figure 7: From top to bottom and left to right: Target hypotheses at scale 1 (original image), target hypotheses at scale 2 (2x downsampled), target hypotheses at scale 3(3x downsampled), original image, and generated target likelihood map. Masks for each scale are resized for visualization.

average of local maxima. Obviously, the critical part is to obtain the appropriate weights. To calculate the weights, first, the local maxima are detected. Then, they are sorted in descending order to form a vector ($LocalMax_{sorted}$). The weights are obtained by calculating the normalized laplacian of this vector since higher laplacian represents distinctive falls. This methodology fits well to our problem since distinctive falls indicate splits between different target hypothesis groups having similar likelihood values; so it achieves successful separation of distinctively more significant target hypotheses. The formal definition of the weighting procedure is shown in Eqn. 4 and 5.

$$Thr = LocalMax^{T}_{sorted}.\nabla^{2}_{norm}\left(LocalMax_{sorted}\right), \quad (4)$$

$$\nabla^{2}_{norm}(f) = \frac{\nabla^{2}(f) - min\left(\nabla^{2}(f)\right)}{\sum_i \nabla^{2}(f)|_i - min\left(\nabla^{2}(f)\right)}. \quad (5)$$

After selection of the target hypotheses as relevant targets, the tracking is simply achieved by matching the relevant targets from consecutive frames just by minimizing spatial distance.

3 EXPERIMENTS

The proposed method was tested for two different aspects: Detection and tracking capabilities. For the detection part, success is defined as detecting all true targets while rejecting non-target background clutter. Thus, to examine the detection performance, two success measures, which are false discovery rate (Eqn. 6) and true positive (Eqn. 7) rate, are used together.

$$FDR = \frac{FP}{FP+TP}, \quad (6)$$

$$TPR = \frac{TP}{TP+FN}. \quad (7)$$

Another important task that should be achieved is tracking of the detected targets. Despite existence of multiple targets in each scenario of the VIVID dataset (VIVID, 2005), the ground truth is only provided for the primary target. Due to lack of ground truth data for secondary targets, we followed the same procedure used in (Bolme et al., 2010). Thus, the tracking performance of the proposed method was evaluated by manually labeling the results as good tracking; tracking had drifted off center, or lost. A track is described as good track when the track center is within the object; labeled as drifted track when the track center is located outside of the object boundary and a track stated to be lost whenever track gate ceases its existence in the presence of the target. One exemplary illustration is given in Fig. 8 for good and drifted tracks respectively.

For the experiments, the VIVID dataset is preferred due to the challenges on each scenario including out of plane rotation, pose variation, occlusion, low contrast, existence of similar targets in the vicinity and defocusing. Since the algorithm is designed to be used with single band images (especially for IR),

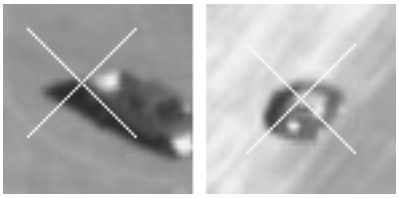

Figure 8: Examplary outputs showing the drifted track, on left, and successful track, on right.

Figure 9: Sample result on VIVID dataset. Top row, columns 1-2: Scale changes. Top row, columns 2-4: Defocusing. Bottom row, columns 1-2: Different motion patterns, changing number of targets. Bottom row, columns 3-4: Occlusion.

the 3-channel sequences of the dataset are converted to grayscale. However, extending the scheme to RGB requires nothing but replacing edge detection phase with an RGB compatible version.

For each scenario, effective saliency and compactness thresholds were set to 0.7 and 0.4, respectively. The variance of the normal distribution that was used to generate target likelihood maps was set to 0.15 and a three-level pyramid structure was used: 1^{st} level processing original image, 2^{nd} level processing original image downsampled by 2 and 3^{rd} level by 3. Since optimum number of scales depends on the span of expected target size, minimum number of scales should manually be selected considering the application. Likewise, shape and window size of the morphological operator should also be selected accordingly. In the testing procedure, a 5x5 circular shaped operator is used.

In Figure 9, some of the important findings of the experimental results are demonstrated. The first two images of the first row illustrates the success of the algorithm against scale changes which is achieved with the usage of pyramid structure. Remaining images of the first row demonstrates the behavior of the proposed method in case of missing observations. In this scenario, the target detection fails for a while due to defocusing of the camera. Despite the missing observations, tracks are continued without breaking and the targets are again well localized after refocusing of the camera. However, one should note that a false alarm is generated after the defocusing since the Canny threshold is automatically adjusted to tolerate the low contrast. The importance of the selection of a simple motion model (maximum velocity) is illustrated on the first two images of the 2^{nd} row. If a restrictive probabilistic motion model was used, some of the targets having different turning angles would be lost. Moreover, these sub-figures also visualizes the success in handling varying number of targets. Finally, last two images of the second row visualizes the major weakness of the proposed algorithm which is the incapability of occlusion handling resulting in track losses.

Table 1: Performance results of proposed method for detection and tracking on VIVID dataset (in percentage %).

Dataset	*False Detection Rate (FDR)*	*True Positive Rate (TPR)*	*Track Quality*
egTest01	0.626	94.098	98.651
egTest02	10.521	77.188	99.440
egTest03	18.127	74.124	88.631
egTest04	4.676	83.203	81.478
Total	8.487	82.153	92.050

The results of the experiments are summarized in Table 1. According to the results, target selection scheme can detect $\approx 82\%$ of the targets with an acceptable FDR of $\approx 8.5\%$. Moreover, $\approx 92\%$ of the detected targets were tracked successfully, meaning that the window was not drifted off of the center of the object. In addition to the detection and tracking performance of the proposed method, another important aspect is the computational load. The proposed solution was tested using un-optimized C++ code running on a single core of an Intel i5-3470 3.2GHz CPU and was able to run at a minimum rate of 30.12fps and an average rate of 35.63fps for maximum 256 target hypotheses at each scale of the pyramid. Note that, the frame-rate can further be improved by using parallel processing or advanced optimization techniques.

The results show that it is possible to have both detection and tracking with a sufficient quality and low computational cost using the proposed method. More importantly, the results imply that it is possible to achieve an acceptable tracking performance by simply using spatial distance minimization of mea-

surements with an appropriate detection scheme.

4 CONCLUSIONS

In this study, a multi-target detection and tracking method designed for real-time systems is introduced. The experiments showed that the proposed algorithm achieves a sufficient true positive rate with a relatively low false discovery rate on the utilized test sets. Moreover, it is also seen that, usage of a successful detection scheme reduces the complexity of tracker; and even with the simplest association scheme, a sufficient tracking performance can be obtained.

Usage of the designed algorithm introduces many advantages including time efficiency, scale-invariance and adaptability to changing number of targets in the scene. Moreover, the algorithm requires no supervision which makes it a suitable option for electro-optical surveillance and reconnaissance systems. On the other hand, the algorithm is shown to have some disadvantages. Although the proposed method can achieve tracking with high frame rates, it has no mechanism for occlusion handling which decreases the performance. Another significant disadvantage of this algorithm is caused by the target hypothesis generation method: Canny edge detection method may fail on low contrast scenes despite its dynamic thresholding scheme since edge detection may fail in low contrast.

As a future work, we plan to employ tracklet concept to increase the performance of the proposed method on the scenes where frequent occlusions are present. Also we plan to work on the target hypotheses generation scheme to make the proposed method invariant to the properties of the input imaging system yielding increased robustness and reliability.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604.

Andriyenko, A. and Schindler, K. (2011). Multi-target tracking by continuous energy minimization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1265–1272. IEEE.

Aytekin, C., Tunalı, E., and Öz, S. (2014). Fast semi-automatic target initialization based on visual saliency for airborne thermal imagery. In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, Visapp'14, pages 490–497.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Berclaz, J., Fleuret, F., Turetken, E., and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1806–1819.

Bolme, D. S., Beveridge, J. R., Draper, B. A., and Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2544–2550.

Canny, J. (1986). A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698.

Cheng, M., Zhang, G., Mitra, N. J., Huang, X., and Hu, S. (2011). Global contrast based salient region detection. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 409–416. IEEE Computer Society.

Fortmann, T. E., Bar-Shalom, Y., and Scheffe, M. (1980). Multi-target tracking using joint probabilistic data association. In *Decision and Control including the Symposium on Adaptive Processes, 1980 19th IEEE Conference on*, volume 19, pages 807–812. IEEE.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Niedfeldt, P. C. and Beard, R. W. (2014). Multiple target tracking using recursive ransac. In *American Control Conference (ACC), 2014*, pages 3393–3398. IEEE.

Otsu, N. (1979). A threshold selection method from gray-level histogram. *IEEE Transactions on System Man Cybernetics*, SMC-9, No:1:62–66.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076.

Prewitt, J. M. (1970). Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19.

Ramakoti, N., Vinay, A., and Jatoth, R. K. (2009). Particle swarm optimization aided kalman filter for object tracking. In *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*, pages 531–533. IEEE.

Reid, D. B. (1979). An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843–854.

Ristic, B., Arulampalam, S., and Gordon, N. (2004). *Beyond the Kalman filter: Particle filters for tracking applications*, volume 685. Artech house Boston.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In *Computer Vision–ECCV 2006*, pages 430–443. Springer.

Sobel, I. and Feldman, G. (1968). A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, pages 271–272.

Tsai, C., Dutoit, X., Song, K., Van Brussel, H., and Nuttin, M. (2010). Robust face tracking control of a mobile robot using self-tuning kalman filter and echo state network. *Asian Journal of Control*, 12(4):488–509.

VIVID (2005). http://vision.cse.psu.edu/data/vivideval/datasets/datasets.html.

Wei, Y., Wen, F., Zhu, W., and Sun, J. (2012). Geodesic saliency using background priors. In *Proceedings of the 12th European conference on Computer Vision - Volume Part III*, ECCV'12, pages 29–2, Berlin, Heidelberg. Springer-Verlag.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13.