Optical Engineering

SPIEDigitalLibrary.org/oe

Salient point region covariance descriptor for target tracking

Serdar Cakir
Tayfun Aytaç
Alper Yildirim
Soosan Beheshti
Ö. Nezih Gerek
A. Enis Cetin

Salient point region covariance descriptor for target tracking

Serdar Cakir
TÜBİTAK BİLGEM İLTAREN
Şehit Mu. Yzb. İlhan Tan Kışlası
2432. cad., 2489. sok.
TR-06800, Ümitköy, Ankara, Turkey
and
Bilkent University
Department of Electrical and Electronics Engineering
TR-06800, Ankara, Turkey
E-mail: serdar.cakir@tubitak.gov.tr

Tayfun Aytaç
Alper Yildirim
TÜBİTAK BİLGEM İLTAREN
Şehit Mu. Yzb. İlhan Tan Kışlası
2432. cad., 2489. sok.
TR-06800, Ümitköy, Ankara, Turkey

Soosan Beheshti
Ryerson University
Department of Electrical and Computer Engineering
Toronto, Ontario, Canada

Ö. Nezih Gerek
Anadolu University
Department of Electrical and Electronics Engineering
İki Eylül Kampüsü
TR-26470, Eskişehir, Turkey

A. Enis Cetin
Bilkent University
Department of Electrical and Electronics Engineering
TR-06800, Ankara, Turkey

Abstract. Features extracted at salient points are used to construct a region covariance descriptor (RCD) for target tracking. In the classical approach, the RCD is computed by using the features at each pixel location, which increases the computational cost in many cases. This approach is redundant because image statistics do not change significantly between neighboring image pixels. Furthermore, this redundancy may decrease tracking accuracy while tracking large targets because statistics of flat regions dominate region covariance matrix. In the proposed approach, salient points are extracted via the Shi and Tomasi's minimum eigenvalue method over a Hessian matrix, and the RCD features extracted only at these salient points are used in target tracking. Experimental results indicate that the salient point RCD scheme provides comparable and even better tracking results compared to a classical RCD-based approach, scale-invariant feature transform, and speeded-up robust features-based trackers while providing a computationally more efficient structure. [DOI: 10.1117/1.OE.52.2.027207]

Subject terms: salient points; feature selection; feature extraction; region covariance descriptor; covariance tracker.

Paper 121317 received Sep. 12, 2012; revised manuscript received Jan. 24, 2013; accepted for publication Jan. 25, 2013; published online Feb. 22, 2013.

1 Introduction

In target tracking, it is important to extract features from the target region that have high differentiation property and scale and rotation invariance. Features should be robust to noise, partially invariant to affine transformation, intensity changes, and occlusion.[1,2] Another issue in target tracking is to estimate and predict target location in the subsequent frames based on the observations.[3] A fundamentally important requirement comes from video processing. In order to process video frames while preserving real-time requirements, it is important to extract features in a computationally efficient manner for object tracking purposes.[4] Features may be the color, raw pixel intensities or statistics extracted from these values, edges, displacement vectors in optic flow-based approaches, textures, and their combinations depending on the target model (appearance and motion) and imaging system. A detailed evaluation of point-of-interest detectors and feature descriptors for visual tracking can be found in Refs. 5 and 6.

Features obtained by scale-invariant feature transform (SIFT)[7] are independent of scale, rotation, and intensity change and robust against affine transformation. As a feature detector, SIFT uses difference of Gaussians. SIFT is widely used in applications for target detection,[8,9] tracking,[9,10] classification,[11] image matching,[12–14] and constructing mosaic images.[15] When compared to other point-of-interest detectors such as Moravec[16] and Harris,[17] SIFT features are more robust to background clutter, noise, and occlusion. Unfortunately, despite the distinctive properties of SIFT, the feature extraction process is time-consuming, and the method is hardly used in real-time applications. Inspired by the previous feature descriptor schemes, the authors of speeded-up robust features (SURF) descriptors claimed that the SURF scheme approximates even outperforms previously published techniques in a more computationally

0091-3286/2013/$25.00

efficient manner.[18] In SURF, the detector is based on the efficient computation of a Hessian matrix at different scales. There are other feature descriptors such features from accelerated segment test,[19] keypoint classification with randomized trees[20] and ferns.[21] A detailed performance comparison of the above-mentioned methods is provided in Ref. 6 for a common database.

The covariance descriptor proposed in Ref. 22 provides an efficient signature set in object detection and classification problems and the descriptor is successfully used in applications, such as indoor and outdoor target tracking,[23] fire and flame detection,[24] sea-surface and aerial target tracking,[25] pedestrian detection,[26] and face recognition.[27]

In our earlier work,[25] we proposed an offline feature selection and evaluation mechanism for robust visual tracking of sea-surface and aerial targets based on region covariance descriptor (RCD). In the feature extraction phase, features were constructed via the RCD, and feature sets resulting in the best target/background classification were used for tracking. The same feature set is used in Ref. 28 for performance comparison of classifiers for maritime applications. The previously proposed target tracking scheme[25] outperformed correlation,[29] Kanade-Lucas-Tomasi (KLT) [30–32] feature, and SIFT-based[7] trackers in both air and sea surveillance scenarios. In that work, gradient-based features, together with the pixel locations and intensity values were observed to be the most powerful features. However, the proposed tracking scheme needs to be significantly accelerated for real time applications. The main reason for the high computation cost is the requirement of extraction of features from all pixels in the target region and the accompanying rules for target update strategy, which takes into account scale changes in different search regions. Motivated by these observations, a computationally efficient technique is proposed for the calculation of the RCD. This alternative descriptor is named salient point region covariance descriptor (SPRCD), and the descriptor provides a computationally efficient approach without losing the classical RCD's representative power. We compared the performance of the SPRCD with the classical RCD-based approach[25] and SIFT-[7] and SURF-based[18] trackers.

In the literature, various researchers have attempted to develop algorithms in order to construct RCD in an efficient way.[22–34] The "integral image" concept is proposed in Ref. 22 to construct RCD in a computationally efficient manner. The region codifference method[33,34] enables further reduction in the computational complexity of the RCD by replacing the multiplication operators with an addition/subtraction-based operator. The covariance descriptor within visually salient regions is computed in Ref. 35 for duplicated image and video copy detection. In the paper, the authors use a maximization type of information theoretic approach to calculate visual saliency maps by employing a data-independent Hadamard transform. Then, they calculate the RCD using the features extracted from local windows centered at the pixels that provide saliency scores exceeding a predefined threshold. In Ref. 36, the subsets of the image feature space are used together with the means of the image features in a computationally efficient manner for human detection problem. In Ref. 37, the characteristics of the eigenvalues of the weighted covariance matrix are used for the position correction task. The weighted covariance matrix proposed in that work is based on the pixel-wise intensity statistics of the reference image and the scene image. The eigenvalues of this matrix are analyzed to determine whether the pixel contains detailed information. Although this technique is not an RCD type of scheme, the local complexity is taken into account to relate the local information with target characteristics. To the best of our knowledge, no attempts for computing RCD at salient points have been made previously for target tracking purposes. In this paper, we propose the utilization of salient points and the RCD approach together to develop a computationally efficient descriptor scheme for target tracking. We investigate the relation between the RCDs computed at each and every pixel and at only salient points and observe that RCD computation can be decreased when the pixel characteristics are taken into account before covariance computation, i.e., the autocorrelation of the pixel with its neighborhood.

The paper is organized as follows: In Sec. 2, SPRCD is briefly described. Feature selection for the descriptor calculation is explained in Sec. 3. In Sec. 4, the target tracking framework is briefly described. Experimental work and results including the performance comparisons over different performance measures, including target loss indications, are provided in Sec. 5. Concluding remarks are presented and direction for future research is provided in Sec. 6.

2 Salient Point Region Covariance Descriptor

The RCD is widely used in various image representation problems due to its low computational complexity and robustness to partial occlusion. It also enables one to add or remove features in a simple manner to adapt the tracker for different target types and imaging systems. However, the cost of computing RCD significantly increases as the image region used for the descriptor calculation grows. This is especially the case when large targets need to be tracked. In order to determine an upper limit to the descriptor computation cost and to satisfy the real-time requirements, the SPRCDs are proposed.

The calculation of the classical RCD starts by stacking the feature matrices $(f_i; i = 1, 2, \ldots, D)$ extracted from an $H \times W$ dimensional image in order to construct $H \times W \times D$ dimensional feature tensor as given in Fig. 1. A detailed discussion for the extraction of feature matrices $(f_i s)$ is provided in Sec. 3. In the feature tensor, the elements in each layer with the index (m, n) are sorted to construct the feature vector $(\underline{S}_t)$ [Eq. (3)]. In the classical RCD, a total of $H \times W$ feature vectors $(\underline{S}_t)$ are constructed:

$$\underline{S}_t = [\, f_1(m,n) \quad f_2(m,n) \quad \cdots \quad f_D(m,n) \,], \tag{1}$$

where $m = 1, 2, \ldots, W$, $n = 1, 2, \ldots, H$, $t = 1, 2, \ldots, k$, and $k = H \times W$.

The computation procedure of the SPRCD is the same as the procedure in classical RCD computation[22,25] up to this point. The main and crucial difference in the calculation of SPRCD is that only the feature vectors corresponding to salient point locations are used instead of using feature vectors at all pixel positions. We tried two different point extractors in the experiments, namely the Harris corner detector[17] and the Shi-Tomasi[32] detector. The covariance descriptors calculated over the corners extracted by the Harris method did not provide satisfactory tracking performances, especially in scenarios where the target template changes

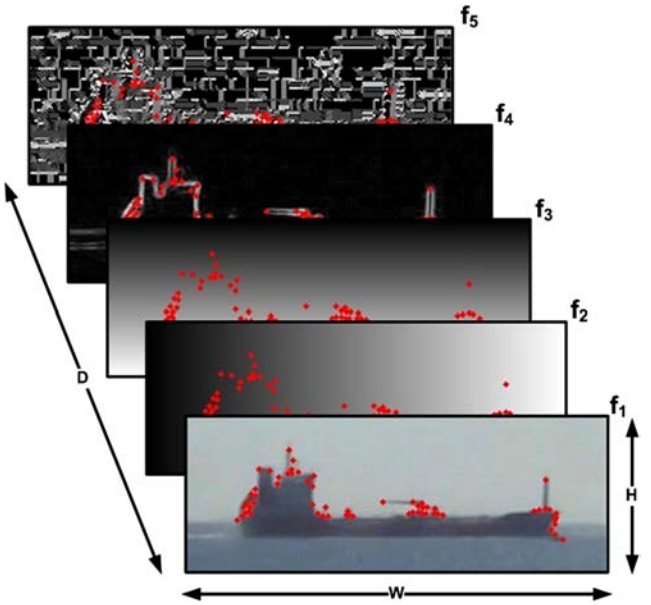

Fig. 1 The illustration of determining salient points in the feature tensor.

rapidly. Therefore, the salient points are determined by the minimum eigenvalue method introduced by Shi and Tomasi. In this method the corner points are determined by analyzing the eigenvalues of the Hessian matrix (H). The method relates the image point characteristics with the values of the two eigenvalues of the matrix H. At this point, instead of recalculating the Hessian matrix directly, the available features used in the SPRCD calculation are gathered in order to construct Hessian matrix. By this way, no additional effort to calculate the Hessian matrix is made. As a reminder, the structure of the Hessian matrix is provided in Eq. (2):

$$H = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix}, \tag{2}$$

where

$$\frac{\partial^2 I}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial I}{\partial x}\right)$$

and

$$\frac{\partial^2 I}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial I}{\partial y}\right)$$

are the second derivatives along the horizontal and vertical axes, respectively and

$$\frac{\partial^2 I}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial I}{\partial y}\right) = \frac{\partial}{\partial y}\left(\frac{\partial I}{\partial x}\right)$$

is the mixed derivative along the horizontal and vertical axes. Two small values of the matrix H mean a roughly constant region, whereas two large eigenvalues indicate a "busy" structure. Such busy regions can correspond to noise, as well as salt and pepper texture, or any pattern that can be tracked reliably.[32] Therefore, a thresholding type of approach onto the minimum eigenvalue of the matrix was developed in Ref. 32 to determine the representative points for tracking.

The main idea behind the descriptor calculation approach using salient points is finding the relational variances between the features located at important corners instead of considering the variances of features calculated at each and every image pixel location. In this way, a representative and computationally efficient feature descriptor is developed. Moreover, the proposed descriptor scheme is not affected by partial occlusion that causes the KLT tracker to fail in target-tracking scenarios.[25] Since the proposed descriptor scheme depends on the spatial relations of the features calculated at corner points rather than a simple corner matching type of approach, it is not affected by the destructive effects of partial occlusion.[25] The illustration utilizing the feature vectors corresponding to the salient points is given in Fig. 1. In Fig. 1, instead of displaying a generic implementation, the depth of the feature tensor is selected as five in order to obtain a reasonable visualization. Suppose that there exists ε salient points extracted within a given region, then the covariance descriptor calculation procedure can be rewritten as

$$M_{\mathrm{SPR}}(p,q) = \frac{1}{\varepsilon - 1}\left[\sum_{t=1}^{\varepsilon} \underline{\mathrm{S}}_t(p)\underline{\mathrm{S}}_t(q) \frac{1}{\varepsilon}\sum_{t=1}^{\varepsilon} \underline{\mathrm{S}}_t(p) \sum_{t=1}^{\varepsilon} \underline{\mathrm{S}}_t(q)\right], \tag{3}$$

where $\underline{\mathrm{S}}_t$, $(t = 1, 2, \ldots, \varepsilon)$ denote feature vectors evaluated only at salient points. Since ε is naturally less than the number of pixels in the target region (k), SPRCD is computationally more efficient than the classical region covariance method. Depending on the scenario, the number of salient points (ε) may vary between tens to hundreds. An upper limit ϖ for ε is determined via extensive experimental work using the relation presented in Eq. (4):

$$\varepsilon = \begin{cases} \varepsilon & \text{if } \varepsilon < \varpi \\ \varpi & \text{if } \varepsilon < \varpi \end{cases}. \tag{4}$$

This strategy prevents the cost of the descriptor complexity from growing limitlessly. In the experiments, the target region is represented with an SPRCD calculated using at most $\varpi = 25$ salient points that provide satisfactory tracking accuracies. Although the upper limit ϖ is selected as 25 after a large-scale experimental framework, it can further be adjusted adaptively by defining a certain ratio between ϖ and the number of image pixels, k.

The RCD can be calculated using the "Integral Image" concept[22] rather than the calculation using the classical formulation [Eq. (3)]. The "Integral Image" method introduces a significant reduction in the computational complexity of RCD. The SPRCD feature extraction scheme proposed herein is implemented over the "Integral Image" concept rather than the classical covariance computation formulation. By this way, a further reduction in the computational complexity is introduced.

In the next section, a brief discussion about the feature set used in the descriptor computation is provided.

3 Feature Selection

The feature set used in SPRCD calculation is determined by using the experimental results obtained in our previous work.[25] The gradient-based feature set $(I, x, y, \mathrm{GM}, \mathrm{GO})$,

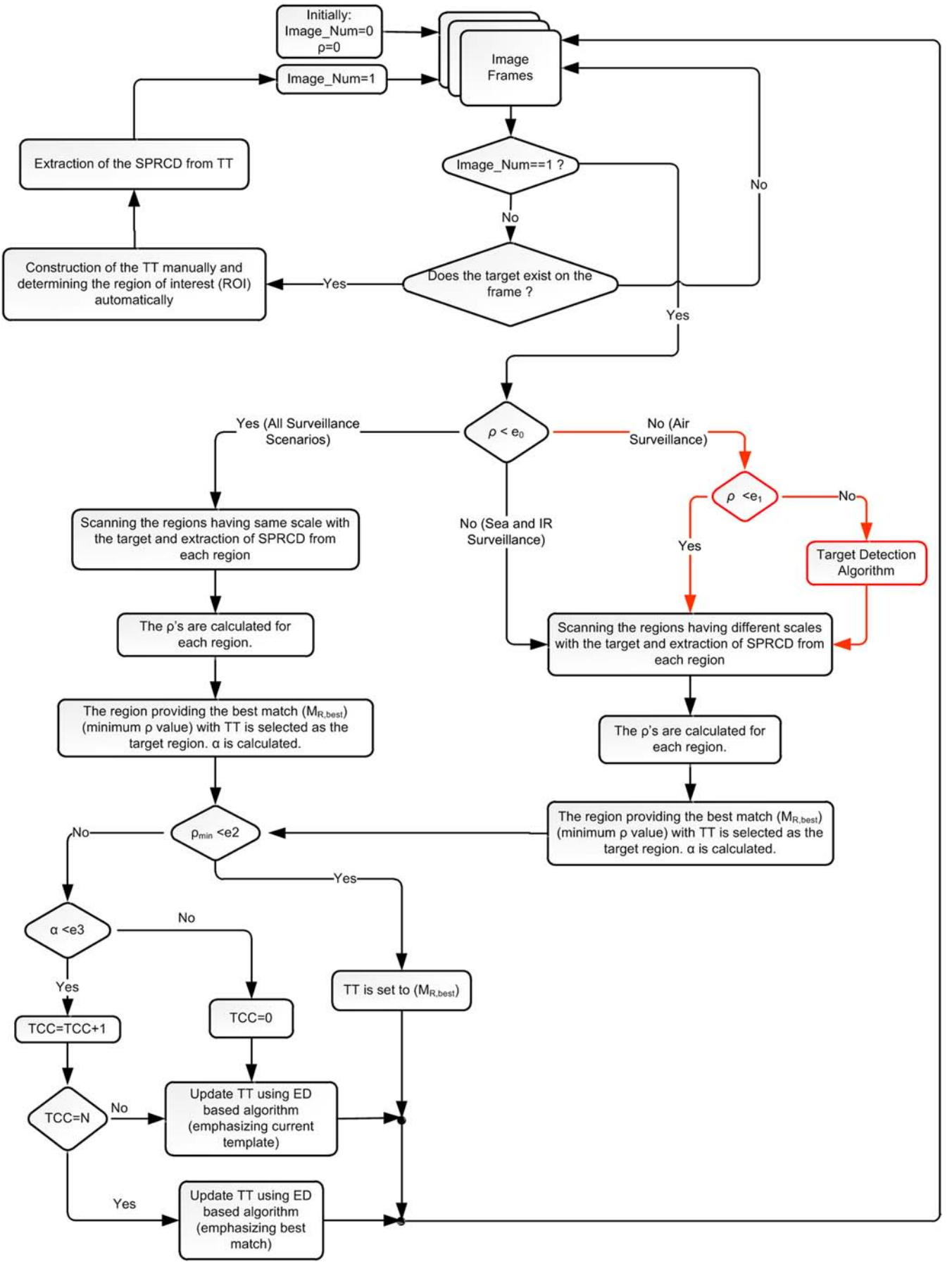

Fig. 2 The flow diagram and TT update strategy of the proposed SPRCD tracker.

which provided plausible and robust tracking results, is used in the feature extraction phase of the proposed descriptor scheme. Here, I denotes the image intensity, x and y denote the horizontal and vertical pixel locations, and GM and GO stand for the gradient magnitude and orientation, respectively. GM and GO features are calculated using the first partial derivatives along the horizontal ($\partial_{1,x} = \frac{\partial I}{\partial x}$) and vertical axis ($\partial_{1,y} = \frac{\partial I}{\partial y}$) as in Eq. (5). It can be noted that the first partial derivatives $\partial_{1,x}$ and $\partial_{1,y}$ are calculated using the filter $[-1,0,1]$.

$$\mathrm{GM} = \sqrt{\partial_{1,x}^2 + \partial_{1,y}^2} \quad \mathrm{GO} = \tan^{-1}\left(\frac{\partial_{1,y}}{\partial_{1,x}}\right). \tag{5}$$

The feature set $(I, x, y, \mathrm{GM}, \mathrm{GO})$ is illustrated in Fig. 1 where f_1, f_2, f_3, f_4, and f_5 denote the features I, x, y, GM, and GO, respectively. All of the features used in the descriptor computations are normalized to $[0,1]$ range.

4 SPRCD-Based Tracker

The general framework of the proposed SPRCD-based tracking scheme is presented in Fig. 2. The proposed tracker is initialized as soon as the target region is determined. After initialization, the determined target gate and the next image frame are exposed to a preprocessing step. The preprocessing step includes deinterlacing and gray-scale conversion for visual band images. In the surveillance applications, the target region is generally determined automatically or manually. In our case, the target region is selected manually by an operator. As soon as the target template (TT) is determined, the target is searched within a search region (SR). The SR is taken as the smallest rectangle surrounding the TT-sized rectangles located at each pixel location within a τ-pixel neighborhood of the target center. At the end, $(2\tau + H) \times (2\tau + W)$ dimensional SR is obtained. The illustration of the SR is given in Fig. 3.

After the determination of the SR, the SPRCD belonging to the TT and the TT-sized subregions within the SR are computed. A descriptor-matching type of approach is performed in order to locate the target in the current frame. In Ref. 22, the descriptor-matching process is carried out by the eigenvalue-based metric defined in Ref. 38. However, in this study, we prefer to use a computationally efficient metric based on normalized L_1 distance[34] presented in Eq. (6):

$$\rho(\hat{M}_{TT}, \hat{M}_R) = \sum_{i=1}^{D}\left\{\sum_{j=1}^{D}\left[\frac{|\hat{M}_{TT}(i,j) - \hat{M}_R(i,j)|}{\hat{M}_{TT}(i,i) + \hat{M}_R(i,i)}\right]\right\}, \tag{6}$$

where $\hat{M}_{TT}$ and $\hat{M}_R$ are the SPRCDs extracted from the TT and the region used for comparison (M_R), respectively.

As visualized in Fig. 2, the tracker algorithm checks the value of ρ to decide which search mode is used in the next video frame. If ρ is larger than a predefined threshold e_0, the target is searched in different scales (meaning camera zoom or target approach/leave). In that case, the SR approach (illustrated in Fig. 3) is modified by increasing or decreasing the target template size rather than fixing it. By this way, different scaled rectangles centered at each pixel of the SR are taken as candidate regions. The dimensions of the different scaled rectangles are determined by multiplying the dimensions of the target template of the previous frame with the scale coefficient κ. The tracker contains two shrinkage ($\kappa = \{0.8, 0.9\}$) and two growth ($\kappa = \{1.1, 1.2\}$) scale coefficients. By this way, the target is searched within the SR using four different scales, considering the target dimension changes in both positive and negative directions. This approach is similar to the Monte Carlo-based target update strategy presented in Ref. 39. The candidate region resulting in the smallest ρ value with the current TT is selected as $M_{R,\mathrm{Best}}$ and the TT is updated using the $M_{R,\mathrm{Best}}$.

In case of scale change, unlike the classical RCD computation, the salient points must be relocated at the scaled TTs. The relocation of salient points is performed using the ratio of the differences between the salient point locations and the location of the center of the TT. The illustration and formulation of the salient point relocation are given in Fig. 4 and Eq. (7), respectively.

$$\begin{aligned} &(p,q) \rightarrow (\tilde{p},\tilde{q}) \\ &\tilde{p} = \tilde{X}_c - \mathrm{sgn}(X_c - p)|X_c - p|\kappa \\ &\tilde{q} = \tilde{Y}_c - \mathrm{sgn}(Y_c - q)|Y_c - q|\kappa. \end{aligned} \tag{7}$$

Here, (p,q) and $(\tilde{p},\tilde{q})$ denote the locations of a certain salient point and corresponding relocated salient point, respectively. Also note that, (X_c, Y_c) and $(\tilde{X}_c, \tilde{Y}_c)$ correspond to the center locations of the TT and scaled TT.

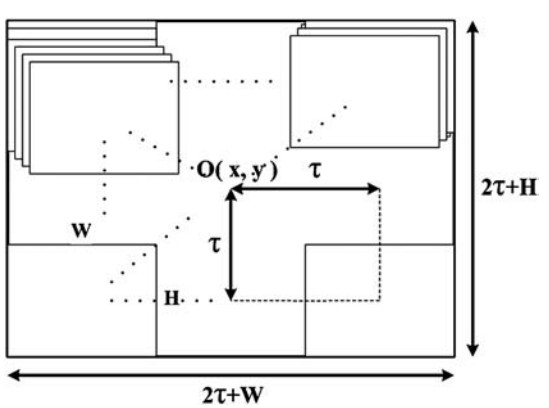

Fig. 3 The illustration of the search region SR. $O(x,y)$ is the target center and W and H are target width and height, respectively.

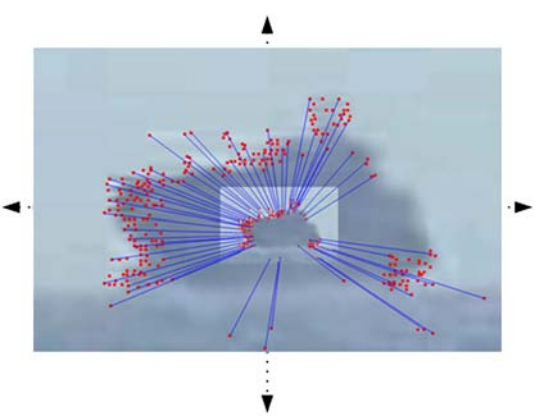

Fig. 4 The illustration of relocation of the salient points in case of scale change. The illustration is exaggerated ($\kappa = 4$) for better visualization of the relocation structure.

After the determination of $M_{R,\text{Best}}$, the TT is updated using a strategy based on the ρ and Euclid distance-based measure (α) defined in Eq. (8):

$$\alpha = \frac{\|M_{R,Best} - TT\|_2}{\text{number of pixels}(M_{R,\text{Best}})}. \tag{8}$$

As can be seen from Fig. 2, the ρ and α terms are used together with their predefined thresholds e_2 and e_3 in the TT update mechanism. If ρ is smaller than e_2, a strong match criterion is satisfied and TT is taken directly as $M_{R,\text{Best}}$. Otherwise, the TT is updated according to the α value. In this case, template change counter (TCC), which is defined to indicate the number of similar ($\alpha < e_3$) TT's and $M_{R,\text{Best}}$'s in the consecutive frames, is altered. If the α value defined in Eq. (8) is less than e_3, TCC value is incremented by one and TT is updated according to Eq. (9):

$$\text{TT}_{\text{Next}} = \alpha(M_{R,\text{Best}}) + (1-\alpha)\text{TT}. \tag{9}$$

In Eq. (9), since α has small values, the previous TT value is more emphasized in the updated TT.

When the TCC reaches a predefined value (N), existing TT is updated with the same strategy, but the $M_{R,\text{Best}}$ is more emphasized in TT update. Therefore, the update in Eq. (9) is modified as follows:

$$\text{TT}_{\text{Next}} = (1-\alpha)M_{R,\text{Best}} + \alpha\text{TT}. \tag{10}$$

In this case, after TT is updated, TCC is reset to zero. The same zero-resetting is also applied if the α value is larger than the threshold e_3.

In the SPRCD-based tracker framework, if TT is significantly different from the $M_{R,\text{Best}}$, the value of ρ becomes greater than its value in a normal match. In this case, the algorithm assumes that the target faced a scale change and initiates a target search with varying scales. This property enables it to track targets with varying scale and shape. It also provides robustness to abrupt camera movements, camera vibrations, and sudden displacements.

In the aerial target tracking case, if ρ is larger than threshold e_1, the tracker assumes that there is a significant change in the target model, and a target detection strategy is initiated in order to adapt the TT to the rapid changes in the target model. The target detection algorithm used in the air surveillance case is a simple intensity thresholding-based technique that takes advantage of contrast difference between the aerial target and the sky background. The reason to use a simple target detection algorithm is to meet the real-time requirements. The detection algorithm is tested over plenty of air surveillance videos, and satisfactory detection performances are achieved.

To sum up, the main difference between the proposed tracking scheme and the one in Ref. 25 is their feature extraction structure. The proposed SPRCD enables a computationally more efficient feature extraction mechanism without losing the representability of the classical RCD.

5 Experimental Work and Results

In the experiments, the proposed SPRCD-based tracker is tested in different scenarios. In this paper, tracking scenarios including sea-surface and aerial targets captured using a visual band camera and a ground target captured using an infrared (IR) camera are provided. The tracking results obtained by the proposed scheme are compared with the tracker structure developed in Ref. 25 that is known to be outperforming the classical tracking algorithms including correlation, KLT, and SIFT-based trackers after performing a large-scale experimental verification. Also the proposed tracking scheme is compared with SIFT- and SURF-based tracking techniques[40] in an appropriate tracking scenario.

The SPRCD-based tracker has naturally different tracking parameters than the classical RCD-based tracker. Since SPRCD structure depends on fewer pixel-wise features, it becomes more sensitive to the changes in the target model. Therefore, the threshold e_0 regarding the descriptor matching result (ρ) must be selected larger than the one used in the classical RCD based structure.

In Sec. 5.1, the performance measures to evaluate the tracking performance are mentioned, and in Sec. 5.2, the tracking results for each tracking scenario are presented.

5.1 *Performance Measures*

In order to evaluate the tracking performance within a quantitative manner, four different morphological similarity measures (PM_i, where $i = 1, 2, 3, 4$) proposed in Ref. 25 are used. The PM_1 and PM_2 are pixel-wise overlapping and nonoverlapping area-based measures, and PM_3 and PM_4 are L_2 and L_1 norms, respectively. A more detailed analysis of these measures as well as a naive performance measure fusion strategy are provided in Ref. 25. By using these performance measures and fusion mechanism, a final evaluation of the tracking performance is established.

In addition to the PM_is, a statistical method based on a confidence interval type of approach[41] is proposed for target loss detection. The target loss detection algorithm is based on an object signature function $[g(z, v)]$ that is the observations of a random variable V with a finite variance. Here, v is the sample of this random variable for any possible values of z. The mean value ($E\{g(z, V)\} = \Gamma(z)$) and the variance ($\text{Var}\{g(z, V)\}$) of the target signature function are used in order to obtain proper confidence intervals with a certain high probability since the standard deviation of the signature function is naturally less than the mean value of the function. The mean value of the signature function is the cumulative distribution function (CDF) of the function and the CDF and variance of the signature function can be estimated using the target parameters of the previous frame. By this way, a target loss detection evaluation mechanism for the current processed image frame can be determined using the mean and variance-based confidence intervals. Let $\Gamma(z)$ denote the mean values of the target signature function where $z = 0, 1, \ldots, 255$ is the value set that a pixel can possess. Therefore, a lower bound $L(z)$ and an upper bound $U(z)$ can be determined as in Eq. (11) around the mean $\Gamma(z)$ by using the Gaussianity assumption for the target signature function due to the central limit theorem:[41]

$$\begin{aligned} L(z) &= \Gamma(z) - \lambda\sqrt{\text{Var}\{g(z, V)\}} \\ U(z) &= \Gamma(z) + \lambda\sqrt{\text{Var}\{g(z, V)\}}. \end{aligned} \tag{11}$$

The parameter λ in the $L(z)$ and $U(z)$ is determined according to the three-sigma (empirical) rule and six-sigma approach. Consequently, $3 \le \lambda \le 5$ becomes a proper

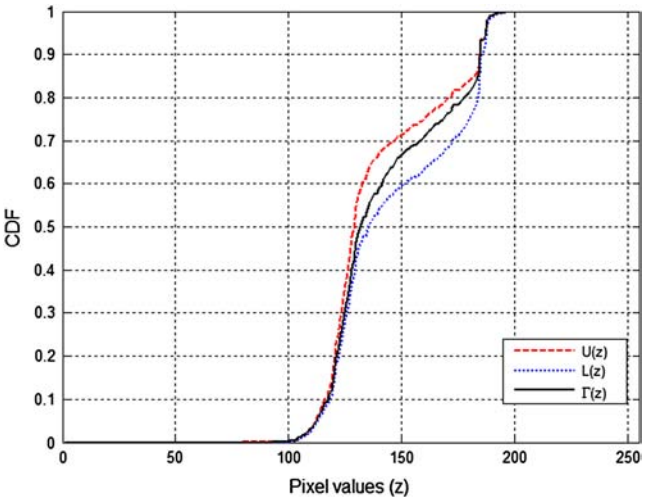

Fig. 5 The bounds on $g(z, V)$ when $\lambda = 3$.

interval for target loss detection problem. As an example, the bounds on $g(z, V)$ using the three-sigma rule ($\lambda = 3$) for a sea-surface target are illustrated in Fig. 5. Note that the bounds on $g(z, V)$ for aerial and IR targets are determined via the same three-sigma approach.

In the experimental results, the average calculation times for RCD and SPRCD blocks and the overall method are also provided. The average processing times for both of the blocks are obtained by averaging the total sum of elapsed times at each visit to the unoptimized descriptor computation block. The proposed tracker is implemented using C++ programming language on a computer with a Core(TM)2 Quad CPU of 2.5 GHz and 2 GB RAM running on Microsoft Windows XP operating system.

5.2 *Tracking Scenarios*

In the first experiment, the RCD- and SPRCD-based trackers are tested in a sea surveillance scenario. The experiment is carried out using a visual band camera that captures 640×480 ($H \times W$) interlaced video frames. In the preprocessing step, a "line doubling" type of approach is used for deinterlacing, where the odd-numbered (even-numbered) rows of each frame are taken and the interpolation of two consecutive rows are placed between these rows. At the end, a reasonably deinterlaced video frame at the same dimension with the original video frame is obtained. The video contains 1000 frames of a moving sea-surface target. The target is occluded by other target-like structures, such as a speed boat and a sail boat. The speed boat moves faster in front of the target of interest (in frames 1 to 500) and causes the "white cap effect" (sea foam) that changes the target environment and contrast rapidly. The sail boat that has low-intensity pixel values moves to the right of the image and occludes the target in frames 850 to 930. The mast of the sail boat causes a sudden intensity change in the target. Consequently, the white-cap effect and the mast of the sail boat are the potential locations that may contain strong corner locations. The tracker parameters τ, e_0, e_2, e_3, and N for sea surveillance scenario are selected as 7, 1, 0.1, 0.0019, and 10, respectively, which are experimentally obtained considering a wide range of cues for sea scenarios. The evaluation of the tracking performances of the classical RCD-based tracker and proposed SPRCD-based tracker are given in Table 1. In the same table, the average computation time for a descriptor is provided in order to observe the computational efficiency of the proposed SPRCD. As seen in the table, both of the trackers result in similar tracking accuracies. The proposed SPRCD approach is 35% faster than the classical one while preserving the track quality. Sample images of the sea surveillance scenario are provided in Fig. 6. According to the target loss detection measure, only four and five out of 1000 frames are determined as the frames that exhibit target losses for the RCD- and SPRCD-based trackers, respectively.

The aerial surveillance scenario is also considered in the experimental studies. The experiments are carried out using the same capture device mentioned above. The video contains 187 frames of a moving helicopter in a cloudy environment. Moreover, the video was captured on a windy day, causing stabilization problems. Therefore, there are some vibrations and sudden movements that reduce the quality of the captured video and make the target tracking task more complicated. The tracker parameters τ, e_0, e_2, e_3, and N for air surveillance scenario are selected as 8, 1, 0.1, 0.0019, and 3, respectively. The performance of the classical RCD- and proposed SPRCD-based trackers is provided in Table 2. The computation time for RCD and SPRCD block is also presented in order to give an idea about the computational complexity of the approaches. In this case, the target is a point-like structure. Therefore, there exist very few salient points extracted from the target region. Consequently, the SPRCD tracker is not able to outperform the classical RCD tracker. Although the SPRCD tracker has lower PM_i values than the classical RCD tracker, the target is tracked with only four target losses until the end of the video. In the same video, the classical RCD-based tracker has two frames containing target losses. The processing time of the proposed approach is more or less the same as the time of the classical RCD as stated before. It is therefore reasonable to conclude that the proposed SPRCD approach is mostly suitable for large targets where the SPRCD takes the advantage of computational efficiency. The sample images for the tracking of the aerial target are provided in Fig. 7.

The proposed SPRCD-based tracking scheme is also tested in an IR surveillance scenario. The IR video used

Table 1 The performance of trackers in visual sea-surface target tracking scenario.

Tracker type	PM_1	PM_2	PM_3	PM_4	Track score	Track loss	Block computation time (milliseconds)
RCD	0.066	0.908	0.99	1.12	0.8375	4/1000	0.1130
SPRCD	0.021	0.849	0.82	0.94	0.8224	5/1000	0.0737

Fig. 6 The sample images of a sea-surface target tracking scenario.

Table 2 The performance of trackers in visual aerial target tracking scenario.

Tracker type	PM_1	PM_2	PM_3	PM_4	Track score	Track loss	Block computation time (milliseconds)
RCD	0.085	0.666	0.87	1.05	0.5998	2/187	0.0665
SPRCD	0.212	0.434	1.71	2.08	0.3230	4/187	0.0719

in this experiment includes 210 frames of a moving vehicle in a complex background that includes stationary objects, buildings, trees, and moving vehicles, and is captured with a longwave IR camera having a frame size of 320×240. The target is also exposed to partial occlusion in certain frames. The sample frames of the tracking results of the SPRCD-based tracker are presented in Fig. 8. The performance of the proposed SPRCD-based tracking scheme is compared with the classical RCD-based framework. Besides, unlike the tracking scenarios presented above, the IR tracking scenario contains a more detailed analysis by introducing SIFT- and SURF-based trackers to the comparison of the tracking results (Table 3). The comparison with SIFT- and SURF-based trackers are not included in the air and sea surface scenarios because in the feature extraction phase, the scenarios include small targets that yield an insufficient set of features. The insufficient feature set due to small targets may degrade the performance of SIFT- and SURF-based trackers; therefore, for a fair comparison, these results are not provided for sea-surface and aerial target tracking. In the IR tracking scenario, the parameters of SIFT and SURF trackers are determined after performing an experimental framework. For the SIFT-based tracker, the number of octave layers is three, contrast and edge thresholds are 1000. σ is 1. Similarly, for the SURF tracker, the number of octave layers is five, and the threshold for the Hessian matrix is 1. The length of the feature descriptor is 128.

From Table 3, it may be concluded that the SPRCD-based scheme outperforms the classical RCD-, SIFT-, and SURF-based tracking schemes. The classical RCD-, SIFT- and SURF-based tracking techniques fail to track the target when most of the target is occluded by another object in

Fig. 7 The sample images of an aerial target tracking scenario.

certain frames. The occlusion also causes the extraction of the SIFT- and SURF-based features to be blocked over the regions overlapped by the occluding object. The proposed SPRCD can handle such situations by considering the covariance type of relations of the Harris corners. In this way, the weak corners that are not considered as strong SIFT and SURF corners, play an important role in target representation. The classical RCD-based trackers fail to track the target when most of the target region is occluded by another target-like structure in certain frames. The target loss indication algorithm verifies the track fail situation by detecting 27 out of 210 losses in this scenario. However, the SPRCD deals with such types of occlusion by taking advantage of the covariance type of relation between the salient points. In that case, only 11 out of 210 frames are detected as the frames that contain target losses. Moreover, SPRCD enables an efficient implementation by reducing the average time of the descriptor calculation block in the IR surveillance case.

Although, the target loss indication scheme gives track loss decision in certain frames of each surveillance scenarios, the targets continue to be tracked. The target loss indication mechanism, in fact, measures the track quality rather than the losses of the target presence. Sudden changes in the target model, abrupt movements and vibrations on the capturing device may be the main reasons for the low track quality.

As the comparison of the "computational time" experiment, the average execution times for a classical RCD and proposed SPRCD computed over different sized $W \times W$ regions are examined. As can be seen from the Fig. 9, the experiment is carried out by selecting a reference point in a visual band video and $W \times W$ target regions are located at this reference point. At each time, the W value is changed and the corresponding elapsed time is computed for the calculation of a descriptor. The computation times for the RCD and SPRCD corresponding to each computation region is visualized in Fig. 10. Note that, both classical RCD- and proposed SPRCD-based trackers track the $W \times W$ sized targets without any track loss conditions. From Fig. 10, one can conclude that the computation time of the classical RCD grows exponentially as the dimensions of the descriptor calculation region increase. However, the increasing size of the calculation region does not have a significant effect on the computation time of the proposed SPRCD since ϖ is fixed to be at most 25. The upper limit for the number of salient points is determined through experimental studies for each tracking scenario. Obviously, one can determine more salient points depending on the scenario by considering the trade-off between the tracking accuracy and the computational cost. Another concern may be the cost of the initial salient point extraction procedure in the case of tracking larger targets. However, this initial cost is not high compared to the inclusion of all pixels in the descriptor computation in the classical RCD approach. Therefore, the proposed SPRCD is computationally more efficient than the classical RCD,

Fig. 8 The sample images of a ground target tracking scenario in IR band.

Table 3 The performance of trackers in IR ground target tracking scenario.

Tracker type	PM_1	PM_2	PM_3	PM_4	Track score	Track loss	Block computation time (milliseconds)
RCD	0.519	0.621	4.94	5.75	0.245	27/210	0.083
SIFT	0.474	0.664	3.60	4.57	0.309	19/210	4.815
SURF	0.057	0.389	5.85	7.74	0.299	14/210	1.118
SPRCD	0.338	0.895	3.37	3.95	0.556	11/210	0.078

especially when dealing with relatively large objects occupying large regions on the image.

In this work, our main aim is to develop a computationally efficient descriptor extraction scheme. Thus, the salient point extraction scheme is employed to modify the classical RCD technique to keep the computational cost as small as possible. However, for more complicated tracking problems, the proposed point selection mechanism can be further expanded by introducing additional points in the descriptor computation. As an additional design, the salient points are expanded by locating a predetermined sized rectangle at the center of the mass of the salient points. The features located at the points in this rectangle are additionally used in the descriptor computation. Hence, the descriptor calculated over these extended salient points provides better tracking accuracies as well as enabling the characteristic of the smooth regions by introducing a predetermined sized rectangle at the center of the mass of the salient points. Although this extended scheme is computationally more efficient than the classical RCD technique, it does not provide the most economic design in terms of computational cost. Since the main concern addressed in this work is the reduction of the computational cost, only the tracking accuracies obtained via the most computationally efficient scheme are included in Sec. 5.

Fig. 9 The illustration of the $W \times W$ computation region located at a reference point. The values for the target size W is selected as follows: $W = \{5, 8, 10, 12, 16, 20, 30, 40, 50, 60, 80, 100, 125, 150, 200\}$

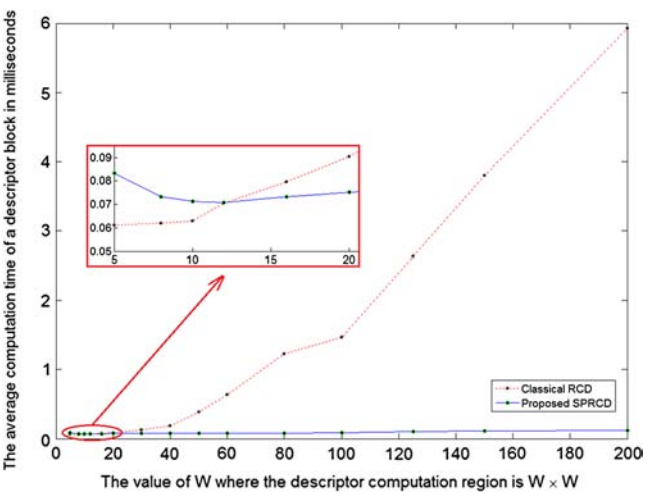

Fig. 10 The computation times of a single classical RCD and proposed SPRCD over the $W \times W$ computation region.

6 Conclusion

In this paper, a new descriptor based on the salient points and RCD is proposed. The proposed descriptor scheme enables robust target tracking as well as computationally efficient structure by using only salient pixels that may have more discriminative power compared to other pixels of a region. The classical RCD has been widely used in many feature extraction problems, but the computational cost of this technique increases excessively when the target region (descriptor calculation region) grows. Hence, the classical RCD scheme may not be implemented in real-time using digital signal processors. By considering only salient points over a region, it is possible to put an upper bound on the computational cost while preserving RCD's power to represent targets. It is experimentally observed that the proposed descriptor even outperforms the classical RCD by using the advantage of variational relations between the salient points in some partial occlusion cases. Moreover, the proposed tracking scheme achieves better tracking accuracies than the well-known SIFT- and SURF-based tracking techniques.

We plan to fuse features obtained using IR cameras operating at different wavelengths and/or visual band cameras. We will investigate the relation of the features at different salient points between images recorded in different bands for robust feature selection. The target loss indication algorithm is intended to be injected into the decision mechanism of the tracker in order to weaken the dependency of the tracker to the direct regional matching metric. In this way, an alternative online control mechanism over the tracker will be introduced.

Acknowledgments

This study is supported by the project with number 109A001 in the framework of TÜBİTAK 1007 Program. The authors would like to thank A. Onur Karali for his efforts in video capture and helpful discussions and Dr. M. Alper Kutay for his support in this study.

References

1. A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Comput. Surveys* **38**(4), 1–45 (2006).
2. H. Yang et al., "Recent advances and trends in visual tracking: a review," *Neurocomputing* **74**(18), 3823–3831 (2011).
3. S. Y. Chen, "Kalman filter for robot vision: a survey," *IEEE Trans. Industrial Electron.* **59**(11), 4409–4420 (2012).
4. X. Zhang et al., "Robust object tracking for resource-limited hardware systems," in *Lecture Notes in Computer Sci., 4th Int. Conf. on Intelligent Robotics and Applications*, H. L. S. Jeschke and D. Schilberg, Eds., Vol. 7102, pp. 85–94, Springer Berlin Heidelberg, Germany (2011).
5. C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. Comput. Vision* **37**(2) 151–172 (2000).
6. S. Gauglitz, T. Höllerer, and M. Turk, "Evaluation of interest point detectors and feature descriptors for visual tracking," *Int. J. Comput. Vision* **94**(3), 335–360 (2011).
7. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
8. C. Park, K. Baea, and J.-H. Jung, "Object recognition in infrared image sequences using scale invariant feature transform," *Proc. SPIE* **6968**, 69681P (2008).
9. T. Can, A. O. Karalı, and T. Aytaç, "Detection and tracking of sea-surface targets in infrared and visual band videos using the bag-of-features technique with scale-invariant feature transform," *Appl. Opt.* **50**(33), 6203–6212 (2011).
10. H. Lee et al., "Scale-invariant object tracking method using strong corners in the scale domain," *Opt. Eng.* **48**(1), 017204 (2010).
11. P. B. W. Schwering et al., "Application of heterogeneous multiple camera system with panoramic capabilities in a harbor environment," *Proc. SPIE* **7481**, 74810C (2009).
12. L. Jing-zheng et al., "Automatic matching of infrared image sequences based on rotation invariant," in *Proc. IEEE Int. Conf. Environmental Sci. Info. Application Technol.*, pp. 365–368, IEEE, China (2009).
13. Y. Pang et al., "Scale invariant image matching using triplewise constraint and weighted voting," *Neurocomputing* **83**, 64–71 (2012).
14. Y. Pang et al., "Fully affine invariant SURF for image matching," *Neurocomputing* **85**, 6–10 (2012).
15. Y. Wang, "Image mosaicking from uncooled thermal IR video captured by a small UAV," in *Proc. IEEE Southwest Sympos. Image Anal. Interpret.*, pp. 161–164, IEEE, New Mexico (2008).
16. H. P. Moravec, "Visual mapping by a robot rover," in *Int. Joint Conf. Artificial Intell.*, pp. 598–600, Morgan Kaufmann Publishers Inc., Japan (1979).
17. C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conf.*, pp. 147–152, University of Sheffield Printing Unit, England (1988).
18. H. Bay et al., "SURF: speeded up robust features," *Comput. Vis. Image Understand.* **110**(3), 346–359 (2008).
19. E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. 9th European Conf. Computer Vision–Volume Part I*, pp. 430–443, Springer-Verlag, Austria (2006).
20. V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(9), 1465–1479 (2006).
21. M. Ozuysal et al., "Fast keypoint recognition using random ferns," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(3), 448–461 (2010).
22. O. Tuzel, F. Porikli, and P. Meer, "Region covariance: a fast descriptor for detection and classification," in *Proc. IEEE European Conf. Computer Vision*, pp. 589–600, Springer-Verlag, Austria (2006).

23. F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on Lie algebra," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recog.* Vol. 1, pp. 728–735, IEEE, New York (2006).
24. Y. H. Habiboğlu, O. Günay, and A. E. Çetin, "Covariance matrix-based fire and flame detection method in video," *Mach. Vis. Appl.* **23**(6), 1103–1113 (2011).
25. S. Cakir et al., "Classifier based offline feature selection and evaluation for visual tracking of sea-surface and aerial targets," *Opt. Eng.* **50**(10), 107205 (2011).
26. S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features," *IEEE Trans. Circ. Syst. Video Technol.* **18**(8), 1140–1151 (2008).
27. Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *IEEE Trans. Circ. Syst. Video Technol.* **18**(7), 989–993 (2008).
28. M. Hartemink, "Robust automatic object detection in a maritime environment: polynomial background estimation and the reduction of false detections by means of classification," Master's Thesis, Delft University of Technology, The Netherlands, Turkey (2012).
29. S. M. A. Bhuiyan, M. S. Alam, and M. Alkanhal, "New two-stage correlation-based approach for target detection and tracking in forward-looking infrared imagery using filters based on extended maximum average correlation height and polynomial distance classifier correlation," *Opt. Eng.* **46**(8), 086401–14 (2007).
30. C. Tomasi and T. Kanade, "Detection and tracking of point features," Technical Report, Carnegie Mellon University (1991).
31. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artificial Intell.*, pp. 674–679, Morgan Kaufmann Publishers Inc., BC, Canada (1981).
32. J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Computer Vision and Pattern Recog.*, pp. 593–600, IEEE, Washington (1994).
33. H. Tuna, İ. Onaran, and A. E. Çetin, "Image description using a multiplier-less operator," *IEEE Signal Process. Lett.* **16**(9), 751–753 (2009).
34. K. Duman, "Methods for target detection in SAR images," Master's Thesis, Bilkent University, Department of Electrical and Electronics Engineering, Ankara, Turkey (2009).
35. L. Zheng et al., "Salient covariance for near-duplicate image and video detection," in *Proc. IEEE Int. Conf. Image Processing*, pp. 2585–2588, IEEE, Belgium (2011).
36. J. Yao and J.-M. Odobez, "Fast human detection from videos using covariance features," in *Proc. European Conf. Computer Vision, Visual Surveillance Workshop*, France (2008).
37. J. Ling et al., "Infrared target tracking with kernal-based performance metric and eigenvalue-based similarity measure," *Appl. Opt.* **46**(16), 3239–3252 (2007).
38. W. Forstner and B. Moonen, "A metric for covariance matrices," Technical Report, Department of Geodesy and Geoinformatics, Stuttgart University (1999).
39. X. Ding et al., "Region covariance based object tracking using Monte Carlo method," in *Proc. IEEE Int. Conf. Control and Automation*, pp. 1802–1805, IEEE, India (2010).
40. A. Vedaldi and B. Fulkerson, *VLFeat: An Open and Portable Library of Computer Vision Algorithms*, http://www.vlfeat.org/ (2008).
41. S. Beheshti et al., "Noise invalidation denoising," *IEEE Trans. Signal Process.* **58**(12), 6007–6016 (2010).

Serdar Cakir received his BSc from Eskişehir Osmangazi University in 2008. Immediately after graduation, he joined Bilkent University and he got his MSc in electrical engineering in 2010. He joined the Scientific and Technological Research Council of Turkey in 2010, where he is currently a research scientist. He also continues his PhD studies at Bilkent University, Department of Electrical Engineering. His main research interests are image/video processing, computer vision, and pattern recognition.

Tayfun Aytaç received his BSc in electrical engineering from Gazi University, Ankara, Turkey, in 2000 and his MS and PhD in electrical engineering from Bilkent University, Ankara, Turkey, in 2002 and 2006, respectively. He joined the Scientific and Technological Research Council of Turkey in 2006, where he is currently a chief research scientist. His current research interests include imaging systems, automatic target recognition, target tracking and classification, and electronic warfare in infrared band.

Alper Yildirim received a BSc degree in electrical engineering from Bilkent University, Ankara, Turkey, in 1996, an MSc degree in digital and computer systems from Tampere University of Technology, Tampere, Finland, in 2001, and a PhD degree in electronics engineering from Ankara University, Ankara, in 2007. He was a design engineer with Nokia Mobile Phones, Tampere. He is currently a chief research scientist with the Scientific and Technological Research Council of Turkey, Ankara. His research interests include digital signal processing, optimization, and radar systems.

Soosan Beheshti received a BSc degree from Isfahan University of Technology, Isfahan, Iran, and MSc and PhD degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2002, respectively, all in electrical engineering. From September 2002 to June 2005, she was a postdoctoral associate and a lecturer at MIT. Since July 2005, she has been with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Ontario, Canada, where she is currently an assistant professor and director of Signal and Information Processing Laboratory. Her research interests include statistical signal processing, hyperspectral imaging, and system dynamics and modeling.

Ö. Nezih Gerek received BSc, MSc, and PhD degrees in electrical engineering from Bilkent University, Ankara, Turkey, in 1991, 1993, and 1998, respectively. During his PhD studies, he spent a semester at the University of Minnesota as an exchange researcher in an NSF project. Following his PhD degree, he spent one year as a research associate at EPFL, Lausanne, Switzerland. Currently, he is a full professor of electrical engineering at Anadolu University, Eskisehir. He is also a member of the Electrical, Electronics and Informatics Research Fund Group of the Scientific and Technological Research Council of Turkey. He is on the editorial board of *Turkish Journal of Electrical Engineering and Computer Science* and *Elsevier: Digital Signal Processing*. His research areas include signal analysis, image processing, and signal coding.

A. Enis Cetin received his PhD from University of Pennsylvania in 1987. Between 1987 and 1989, he was an assistant professor of electrical engineering at University of Toronto. He has been with Bilkent University, Ankara, Turkey, since 1989. He was an associate editor of the *IEEE Transactions on Image Processing* between 1999 and 2003. Currently, he is on the editorial boards of *Signal Processing and Journal of Advances in Signal Processing* and *Journal of Machine Vision and Applications*, Springer. He is a Fellow of IEEE. His research interests include signal and image processing, human-computer interaction using vision and speech, and audiovisual multimedia databases.