Optical Engineering

SPIEDigitalLibrary.org/oe

Region descriptors for automatic classification of small sea targets in infrared video

Martijn M. Mouthaan
Sebastiaan P. van den Broek
Emile A. Hendriks
Piet B. W. Schwering

Region descriptors for automatic classification of small sea targets in infrared video

Martijn M. Moutaen

Defence Materiel Organisation
Directorate of Materiel Royal Netherlands
Navy
Weapon and Communication Systems
v.d Burchlaan 31
NL-2597 PC The Hague, The Netherlands

Sebastiaan P. van den Broek

TNO Defence, Security and Safety
Business Unit Observation Systems
P.O. Box 96864
NL-2509 JG The Hague, The Netherlands

Emile A. Hendriks

Delft University of Technology
Faculty of Electrical Engineering
Computer Vision Lab
Pattern Recognition and Bioinformatics Group
Mekelweg 4
NL-2628 CD Delft, The Netherlands

Piet B. W. Schwering

TNO Defence, Security and Safety
Business Unit Observation Systems
P.O. Box 96864
Oude Waalsdorperweg, 63
NL-2509 JG The Hague, The Netherlands
E-mail: piet.schwering@tno.nl

Abstract. We evaluate the performance of different key-point detectors and region descriptors when used for automatic classification of small sea targets in infrared video. In our earlier research performed on this subject as well as in other literature, many different region descriptors have been proposed. However, it is unclear which methods are most applicable to use on the type of infrared imagery as used onboard naval ships. The key-point detector should detect points of interest that can be used to effectively describe the objects in the imagery. On the basis of the detected key points, the descriptors should discriminate between different classes of small sea targets while being robust to differences in viewing conditions. We propose a similarity measure based on the distance between key-point location and the Euclidean distance between descriptors to quantify the similarity of images. For performance evaluation, we use the receiver operator characteristic as the criterion to rank the evaluated methods. We compare the Harris-, blob- and scale-invariant feature transform (SIFT) detectors and the square neighborhood, steerable filters, invariant moments, and SIFT descriptors. We conclude that the Harris detector combined with the square neighborhood of size 19×19 or the SIFT descriptor results in the best classification performance for our data set.

© 2011 Society of Photo-Optical Instrumentation Engineers (SPIE). [DOI: 10.1117/1.3549907]

Subject terms: classifications; key points; region descriptors; Harris; scale-invariant feature transform; infrared; receiver operator characteristics.

Paper 100802R received Oct. 5, 2010; revised manuscript received Jan. 5, 2011; accepted for publication Jan. 12, 2011; published online Mar. 9, 2011.

1 Introduction

An important aspect of modern naval missions is the defense against asymmetric threats. The classic scenario of a missile attack at high seas is not considered most likely to happen nowadays. The concern about piracy and terrorist attacks is strong and very contemporary. In these scenarios, mainly small boats, such as rigid-hulled inflatable boats (RHIBs), speedboats, and jet skis, are used to approach the target. For naval ships, terrorist attacks are considered the greatest threat. Terrorists are known to use rocket-propelled grenades and improvised explosive devices to damage the ship and harm the crew. To effectively defend the ship against such attacks, it is essential to recognize when an object becomes a threat. For this reason, it is necessary to obtain as much information as possible from all boats in the vicinity of the ship. In this process, information from automated optical and infrared sensors is not yet used to the fullest extent, but this information is very valuable for classification of nearby targets up to 1 km. If it is possible to have automatic classification of targets from infrared imagery, then the decision-making process to determine if an object is a threat can be improved.

In earlier research at the Netherlands Organization for Applied Scientific Research (TNO), several aspects of the use of electro-optical sensors were examined in the research program "Electro-optical sensor systems,"¹ started in 2006. For this program, a database is created with infrared images of cabin boats, RHIBs, and jet skis, among others. These images are representative of the type of imagery available on naval ships. Recently, TNO performed research on automatic classification of small boats in coastal environments² based on this data set.

The goal of the research described here is to evaluate the performance of region descriptors when used for automatic classification of small sea targets in infrared images. Region descriptors are generally used to compare visible-light images with high detail, which makes this not an obvious choice for infrared imagery. This research is an addition to the earlier research, where the usefulness of region descriptors was shown. New methods for the classification process are used for the same data set. The focus of our present research is different because our interest is to evaluate how the classification performance depends on the choice of a key-point detector and region descriptor. The approach of our experiment was inspired by a leading research paper by Mikolajczyk and Schmid³ about a performance

Fig. 1 Processing chain for classification.

evaluation of widely used key-point detectors and region descriptors.

The structure of this paper is as follows. In Sec. 2, we give some information about related work that has been the basis for the research described here. In Sec. 3, we describe the methods used in the classification process and, in Sec. 4, we explain how the performance of different methods is evaluated. In Sec. 5, we present the results of the performance evaluation, and in Sec. 6, the results are discussed and suggestions are made related to the research focus and possible approaches to improve the classification results. Section 7 contains the conclusions.

2 Related Work

In earlier work,² we reported on automatic classification of small boats at sea with different electro-optical systems. In that research, the focus was on the use of global as well as local image features for classification. Furthermore, we addressed the influence of clutter on the classification result and gave a theoretical approach on the misclassification cost, when taking into account that, in practice, there are more classes of boats than defined in the used data set. For the classification based on local image features, they use the Harris corner detector combined with region descriptors based on Gabor filters, invariant moments, and a scale-invariant feature transform (SIFT)-like method, using local image gradients.

In the field of region descriptors, Mikolajczyk and Schmid³ presented valuable work. For example, they developed key-point detectors that are insensitive to affine transformations, with associated affine-invariant region descriptors. They report on a performance evaluation of local descriptors, and different key-point detectors and region descriptors are compared for a large variety of image deformations.³ Their approach of evaluating different combinations of methods has been a guideline for our research.

However, their affine-invariant methods are expected to be less suitable when used with our data set because our data set consists of low-resolution images with low contrast and little texture. Also, the evaluation criterion used is not applicable for our problem. For this, we looked at the approach of Carneiro and Jepson.⁴ They show that their region descriptor, using phase-based local features, performs better than differential invariants.

For classification of ships, generally features describing the extracted shape are used, such as contours,¹³ moments,^{14–16} and Fourier descriptors.¹⁶ Because of the lack of detail on targets in infrared imagery, classification methods based on key points were not used. However, the previous

research² showed that even for small targets, these local descriptors can have better results. For our experiment, we used the results of the earlier TNO research as a comparison for our classification performance. We used the experimental setup as a starting point and improved the classification process by adding or changing some of the steps, based on previously described methods.^{3,4}

3 Classification Process

The classification process is described by the processing chain of Fig. 1. The data set we used for our experiment consists of image patches extracted from infrared video streams of recordings of three types of small boats. The image patches are obtained by using an existing target detection and segmentation algorithm.⁵

The first step in the processing chain is normalization of the data-set images. This is done by stretching the image intensities between 0 and 255 and resizing each image to match the size of the largest images (71×31). To maintain the aspect ratio, the images are zero padded, if necessary. Figure 2 shows a normalized image for every class in the data set.

The emphasis in this research is on using information in small regions on an object for classification, which consists of two steps. Key-point detection determines regions that possibly contain useful information, such as the following:

1. Harris corner detector
2. blob detector, to find localized intensities
3. SIFT key-point detector, which indicates changes in intensity

The next step is describing these regions, to be matched with regions for images in a database. The evaluated region descriptors are based on the following:

1. basic region (the pixel values around the key point)
2. invariant moments, indicating distribution and (a)symmetry of intensity
3. steerable filters, indicating directional information
4. SIFT, describing gradients in intensity around the point at different scales

Matching of two images is done by comparing the key-point locations and region descriptors in both images. Key points are matched if their relative positions are sufficiently close and their descriptors are similar. On the basis of the number of matching key points between a test-set image and the images in the reference set, the test-set image is classified

Fig. 2 Data-set images: side view of (a) cabin boat, (b) RHIB, and (c) jet ski.

by assigning the corresponding class label of the reference image having the highest similarity score (see Sec. 3.3 for details).

Compared to images used in most research papers about region descriptors, we make a few observations. The type and quality of the images in our data set differs considerably from images usually used for comparable experiments, as follows:

1. operational infrared video frames in our data set compared to generated or special images, containing clear characteristics to evaluate applied methods
2. infrared images with a small range of intensities in our data set compared to full color images with a large dynamic range often used for typical data sets
3. low resolution of our images (max 71×31) compared to high-resolution images (800×600 or higher)
4. low contrast and little texture of the objects in our infrared images compared to high contrast and rich textures of full color images

Because of these characteristics of our data set, we expect that the more advanced methods proposed in Mikolajczyk and Schmid³ will not lead to better results than the fundamental methods on which they are based.

3.1 Key-Point Detectors

For our experiment, we selected the fundamental Harris and blob detectors and included SIFT as a more advanced method. The Harris detector⁶ is based on the Hessian or second-moment matrix and detects corners. In the ready-to-use implementation⁷ we used, parameters can be specified for the Gaussian smoothing with standard deviation (σ), the threshold on corner strength, and the region for nonmaximum suppression.

The blob detector⁸ is also based on the Hessian. The parameters specified are the threshold on blob strength, the minimum and maximum size of a blob, and whether to use scale-space estimation. The difference between the Harris and blob detector is determined by the way the Hessian is used. A local maximum of the determinant of the Hessian (H) indicates the presence of a blob structure

$$H = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}, \quad (1)$$

$$\det(H) = L_{xx}L_{yy} - L_{xy}^2, \quad (2)$$

where L_{xx} , L_{yy} , and L_{xy} are partial image derivatives.

A corner is indicated by a local maximum of $f(H)$, where $f(H)$ is defined by a combination of the determinant and the

trace of the Hessian as follows:

$$\begin{aligned} f(H) &= \det(H) - k \text{trace}^2(H) \\ &= (L_{xx}L_{yy} - L_{xy}^2) - k(L_{xx} + L_{yy})^2. \end{aligned} \quad (3)$$

The empirical constant k is kept at the suggested value of 0.04.

The SIFT algorithm^{8,9} is different from the Harris and blob detectors in that it is an inherent affine-invariant key-point descriptor. The algorithm consists of a standard combination of a key-point detector and region descriptor. Key points are detected based on the difference-of-Gaussian (DoG) function, which is much like the multiscale blob detector. From the DoG points, the local maxima across scales are selected. The points that lie on edges are eliminated, and for the resulting points, the local orientation is determined and a gradient magnitude and orientation are assigned to each key point. The orientation assignments are sampled over a region around each key-point location, and orientation histograms are made for 4×4 subregions. Each histogram has eight orientation bins, and the SIFT descriptor is created by putting the histogram values for each subregion into a vector. The length of the descriptor is 128.

3.2 Region Descriptors

The descriptors used in combination with the above-mentioned key-point detectors are square pixel neighborhoods of different sizes, invariant moments, steerable filters, and the SIFT descriptor.

The basic region descriptor consists of the (normalized) values in the pixel neighborhood of the detected key points. The construction of this descriptor is computationally efficient. The classification performance of this descriptor is expected to be worse than the more complex descriptors, such as steerable filters and SIFT. The latter enhance the characteristic features that are present in the regions, making the descriptors more distinctive.

The invariant-moments descriptor we used for this experiment included the scale- and rotation-invariant Hu¹⁰ moments, which are computed for a region of pixels around the detected key-point locations. Different moments indicate different aspects of distribution of intensity, such as symmetry and eccentricity.

The steerable filters descriptor¹¹ consists of the following filter responses: Two filter types, each for six different orientations at three different scales, and two anisotropic filters, the Gaussian and the Laplacian-of-Gaussian. The filters were applied to one of the basic region descriptors to get the filter responses. The resulting descriptor size depends on the filter size. The descriptor is obtained by concatenation of the maximum responses of the four filter types.

Fig. 3 Correction for size of objects because of differences in scale and aspect angle. Scale ratio is determined based on thresholded (binary) images: (a) thresholded reference image and (b) thresholded test image.

For the SIFT descriptor, we used an available ready-to-use implementation¹² of Lowe's algorithm.⁸ Because the original SIFT algorithm uses the DoG key points for the construction of the descriptor, we slightly modified the algorithm to be able to use only the SIFT descriptor part in combination with our other key-point detectors.

3.3 Matching and Classification

Given the selection of key-point detectors and region descriptors, we need to specify how region descriptors from different images are matched. On the basis of the amount of matching key points and the similarity of their region descriptors, the images can be classified as a member of one of the three defined classes: cabin boat, RHIB, or jet ski. In the classification process, we compare only descriptors of the same type. For each descriptor type, we perform a separate classification run.

Matching two images is done in three basic operations. The first step is to compare the width and height of the object in the test image to the reference image to determine the difference in scale/aspect ratio. On the basis of the ratio of the width and height of the object in both images, the test image is scaled to match the object size of the reference image. The key points of the test image are translated to maintain their location on the object.

The second step is to match the key points based on their location on the object. On the basis of the center-of-mass (CoM) of the object in each image, the key-point locations can be specified as vectors with the CoM as the origin. The CoM of the targets is determined based on the thresholded (binary) images, as shown in Fig. 3,

$$\overrightarrow{\text{KP}}_{\text{im}}^i(x, y) = \text{KP}_{\text{im}}^i(x, y) - \text{CoM}_{\text{im}}(x, y),$$

$$i = 1 \dots n \text{ key points}_{\text{im}}. \quad (4)$$

After correction of the key-point vectors of the test image for the scale difference, the distances (pixel units) between the key points are determined. A key point is considered a possible match if the key-point distance is below a specified threshold, τ (see Fig. 4). Without this condition, it is possible to match key points with similar region descriptors while having a different location on the target (e.g., a key point on the corner of the bow could match a key point on the corner of the roof). The use of key-point context is an addition compared to the earlier TNO algorithm and makes the classification process more robust.

$$\left\| \overrightarrow{\text{KP}}_{\text{tst_im}}^i(x, y) - \overrightarrow{\text{KP}}_{\text{ref_im}}^j(x, y) \right\| \leq \tau. \quad (5)$$

The third step in the matching process is to determine how many of the possible matches are correct matches. For this, the region descriptors of the possible matching key points are compared. A possible match is considered a correct match if the Euclidean distance of the region descriptors is below a specified threshold, ε ,

$$D_{\text{euclidean}}^{i,j} = \sqrt{\sum_{n=1}^N \left[\text{Descr}_{\text{tst_im}}^i(n) - \text{Descr}_{\text{ref_im}}^j(n) \right]^2} \leq \varepsilon. \quad (6)$$

where i and j are the possible matching key points to which the descriptors correspond and the summation is over the length of the descriptors, N .

For the classification of images, we specify a similarity measure as shown in Eq. (7). This similarity score is defined as a combination of a key-point score and a descriptor score. The key-point score indicates how many of possible key-point matches were found and is defined by the ratio of matching key points and total number of key points available for matching. Unlike the precision specified by Mikolajczyk and Schmid,³ it allows for comparison between images with different numbers of detected key points. The descriptor score is defined by the mean of the distances of the region descriptors. It is simpler than the similarity score used by Carneiro and Jepson,⁴ but similar to that score, allows for more descriptive key points to have a larger effect on the value,

$$\text{Similarity score} = 100 \% \frac{2 \cdot n \text{ matches}}{n \text{ keypoints}_{\text{im1}} + n \text{ keypoints}_{\text{im2}}} \times \frac{1}{n \text{ matches}} \sum_{m=1}^{n \text{ matches}} \left(1 - \frac{D_{\text{euclidean}}^{i,j}(m)}{\varepsilon} \right). \quad (7)$$

The key-point score is 1 if in both images, an equal amount of key points is detected and all key points are matched. It is lower if either not all possible matches are made or if the numbers of key points differ between the images. The descriptor score is 1 if all distances between the region descriptors are 0. This is only the case when the region descriptors are identical. Multiplication of the key-point score and descriptor score gives a measure for the similarity of the images, based on the key points in each image. Matching identical images gives a similarity score of 100 %.

In the experimental setup, the data set described earlier is used for the performance evaluation of the chosen key-point

Fig. 4 Matched key points (light gray) described as vectors originating in the center-of-mass of objects: (a) reference image and key-point vectors and (b) original test image (not scaled) and key-point vectors.

detectors and region descriptors. Images are split in a reference and a test set. All images from the test set are classified by comparing each image from the test set to all the images from the reference set. As explained earlier, the reference set consists of 10 side-view images and five front-/rear-view images per class. Thus, classification of a test image results in 15 similarity scores for each class, thus, 45 similarity scores for three classes. For each class, the maximum similarity score is used for classification. The test image is eventually assigned to the class that has the highest similarity score. In the next sections, we will refer to this as maximum selection.

4 Evaluation Method

The evaluation method used is based on the receiver-operator characteristic (ROC), and the approach is similar to that used in Carneiro and Jepson.⁴ In our experiment, we vary the threshold on the similarity scores between 0 and 100 % to generate the ROC curves. The true-positive (TP) ratio (TPR) is defined as the proportion of test set images for which the similarity score of the correct class is above the evaluated threshold. The false-positive (FP) ratio (FPR) is the fraction of test set images for which the similarity score of the incorrect class is above the evaluated threshold. For our performance evaluation, we defined two types of ROC curves, the all-class ROC curve and the max-class ROC curve. The all-class ROC curve indicates the class distinctiveness of the region descriptors, while the max-class ROC curve indicates the classification performance.

4.1 All-Class Receiver Operator Characteristic

For the all-class ROC curve, the similarity scores for all the classes (cabin boat, RHIB, jet ski) are evaluated. Every test set image can produce a TP as well as FP, depending on the threshold value that is evaluated. Typical similarity scores for the classes are shown in the first three columns, cl_C , cl_R , and cl_J , in Table 1.

For every evaluated threshold value, the number of true positives and false positives for the consecutive classes are summed to produce the ROC curve. For a threshold value of 100, no TPs and FPs are found, resulting in the (0,0) point in the ROC. While for a threshold value of 0, the similarity scores for one test-set image, result in one TP (see the dark gray cells in Table 1) as well as two FPs (see the light gray cells in Table 1), resulting in the (1,1) point in the ROC.

For an example threshold of 4, the bold similarity scores in Table 1 indicate that the similarity score is above the threshold. The bold similarity scores in the dark gray cells are

the TPs, while the bold similarity scores in the light gray cells are the FPs. Summation of the number of similarity scores >4 , results in eight TPs and four FPs. When we normalize the TPs by the total number of possible TPs, in our example this is 9, the TPR is 0.889. If we do the same for the FPs, then the FPR is 0.222.

The all-class ROC curve is a typical ROC and in our case gives an indication of the distribution of the similarity scores. A steeper ROC curve means that relatively many true positives are detected before most of the false positives are detected. A classification method showing a steep all-class ROC curve is more distinctive than methods with straighter ROC curves.

4.2 Max-Class Receiver Operator Characteristic

Because classification is performed by maximum selection of the similarity scores of the three classes, the all-class ROC cannot be used to indicate the best classification performance. Therefore, we use the max-class ROC. For this ROC, only the maximum similarity score for every test-set image

Table 1 Typical similarity scores.

					Similarity scores (classification result)			
					<i>All-class</i>			<i>Max-class</i>
					cl_C	cl_R	cl_J	$Max(cl_C, cl_R, cl_J)$
Test image indices	1 - 75	C			10.4664	3.2766	3.0493	10.4664
					23.3149	5.0353	2.5631	23.3149
					⋮	⋮	⋮	⋮
					16.6276	3.6586	3.1133	16.6276
	76 - 150	R			3.0916	15.3941	0.0000	15.3941
					0.7071	2.4656	1.6182	2.4656
					⋮	⋮	⋮	⋮
					3.3119	18.6734	1.3706	18.6734
	151 - 225	J			1.7252	2.2883	6.9234	6.9234
					5.4344	9.5109	8.6236	9.5109
					⋮	⋮	⋮	⋮
					2.9339	4.4561	34.8825	34.8825

C = cabin boat class, R = RHIB class, J = jet ski class. Dark gray cells result in true positives when above the evaluated threshold. Light gray cells result in false positives when above the evaluated threshold. Bold scores indicate true positives or false positives for an example threshold on similarity score of 4.

is evaluated. The maximum similarity scores are shown in the rightmost column in Table 1. For the max-class ROC, a test-set image can only produce one true positive or one false positive, if the maximum similarity score is above the evaluated threshold value.

For the example threshold of 4, we now find eight TPs and only one FP. However the maximum number of TPs and FPs is now equal to the number of test-set images. In our example, we only have nine maximum similarity scores, giving a TPR of 0.889 and a FPR of 0.111.

For a threshold value of 100, no TPs and FPs are found, whereas for a threshold value of 0, all the maximum similarity scores result in either a TP or a FP. In this case, all test-set images that are not classified correctly; they are classified incorrectly. Thus,

$$\text{TPR} + \text{FPR} = 1. \quad (8)$$

As a consequence, the max-class ROC curves can never cross the line $\text{TPR} = 1 - \text{FPR}$, where TPR and $\text{FPR} \in [0,1]$ and will never reach the (1,1) point, like a typical ROC.

Despite of this artifact, the max-class ROC curve can be used as a criterion for our performance evaluation. It remains valid that the point on the ROC curve that is closest to (0,1) is the unconditional optimal setting for the method on hand. Especially, the area under the ROC curve (AUC) can be used to distinguish the classification performance of the evaluated methods.

5 Results

The images in the data set are divided into three classes of small boats: cabin boats, RHIBs, and jet skis. Each class has 85 objects with an aspect angle varying from 0 to ± 90 deg, with respect to a left or right side view. An aspect angle between 0 and ± 20 deg is considered a side view, and an aspect angle between ± 60 and 90 deg is considered a front/rear view. The size of the images varies between 26×16 and 71×31 pixels, which correspond to recording distances between 100 and 500 meters.

For the matching and classification process, we created a reference set by selecting five side-view images from each class as ground truth. We also stored mirrored copies of the side views to make the reference-set invariant for the type of side view (left side or right side). Because there is a large difference between side views and front/rear views, we also added five front/rear views to the reference set. Classification of a test image results in 15 similarity scores for each class, thus, 45 similarity scores for three classes.

Classifications runs were performed for different values of parameters in the key-point detection, region description, and matching steps.

For the basic region descriptor, we evaluated neighborhood sizes of 9, 11, 15, 17, 19, and 21. The descriptor dimension depends on the region size and varies between 81 and 441.

The region used to calculate the Hu moments on is the 19×19 neighborhood. Because the Hu moments are defined by seven moments, the descriptor length is 7.

For the steerable filters, we used a filter size of 3 and applied the filters on the 19×19 basic region descriptor, resulting in a 17×17 descriptor. Concatenation of the maximum responses of the four filter types results in a descriptor dimension of $4 \times 17 \times 17 = 1156$.

Fig. 5 Classification performance of pixel neighborhood descriptors.

For each combination of key-point detector and region descriptor, a classification run is performed. We empirically determined suitable ranges for τ and ε . Because we used the Euclidean distance as similarity measure for the descriptors, the average distance for a descriptor type depends on the dimensionality of the descriptor.

We generated ROC curves for all values of τ and ε and all combinations of key-point detector and region descriptor, resulting in a total number of 1080 curves. For the comparison of ROC curves, we used the AUC for $\text{FPR} \leq 5\%$, as a criterion. We split the performance evaluation into three steps: (i) basic region descriptor size, (ii) comparison of key-point detectors and descriptor types, and (iii) aspect angle dependence.

5.1 Basic Region Descriptor Size

To find out which neighborhood size is most suitable for our data set, we first evaluated the classification performance of the basic region descriptors. A small region contains little information and will be less descriptive, while a region that is too large will overlap with other regions or will be partly outside the image boundaries.

The results in Fig. 5 show that, in combination with the Harris detector, the basic region descriptor of size 19×19 gives the best classification performance. This size is related to the used image sizes. Larger regions lose local detail and are hampered by edge effects. The results for combinations with the blob and SIFT detectors are not shown because the Harris detector is the preferred detector. The choice for the Harris detector is explained in Sec. 5.2. For the rest of the experiment, we will only work with the 19×19 basic region descriptor and compare it to the steerable filters, invariant moments, and SIFT descriptors.

5.2 Comparison of Key-Point Detectors and Descriptor Types

We used the area under the ROC curve to find the best result for each combination of key-point detector and region descriptor. As shown in Table 2, the Harris detector performs best for all the region descriptors. In Table 3, the TPRs for

Table 2 Best results for each combination of key-point detector and region descriptor, based on the area under ROC.

AUC max-class ROC for FPR ≤ 0.05			
Descriptor	Harris	Blob	SIFT
9×9 region	0.0271	0.0179	0.0210
11×11 region	0.0319	0.0201	0.0169
15×15 region	0.0350	0.0250	0.0222
19×19 region	0.0422	0.0225	0.0212
21×21 region	0.0378	0.0228	0.0190
Steerable filters	0.0385	0.0245	0.0213
Invariant moments	0.0218	0.0067	0.0136
SIFT	0.0395	0.0297	0.0125

all descriptor types combined with the Harris detector are shown.

We can see that the SIFT descriptor and the 19×19 basic region descriptor have the highest TPR of 92% for a FPR of 5%, followed by the steerable filters descriptor with a TPR of 88%. However, we also note that the AUC for the 19×19 region is largest. This becomes clear in Fig. 6(a).

For a FPR of 5% the TPR of both the SIFT descriptor and the 19×19 descriptor are equal, but for a FPR of 0%, the 19×19 descriptor has a TPR of almost 70%. For the SIFT descriptor, the TPR is approximately 10% lower and for the steerable filters descriptor, the TPR is even 30% lower. For FPRs between 0% and 5%, most of the max-class ROC curve for the SIFT and steerable filters' descriptor is below the 19×19 curve, which explains the higher AUC for the 19×19 descriptor. Hence, for our data set,

the Harris-(19×19) method performs best, followed by the Harris-SIFT method and the Harris-steerable filters method.

In Fig. 6(b), we can see that the Harris-SIFT method is more distinctive than the Harris-(19×19) method, as the Harris-SIFT curve is steeper than the Harris-(19×19) curve. Nevertheless, the classification performance of Harris-SIFT is worse, indicated by the lower AUC of the max-class ROC.

Thus, despite being more distinctive, the Harris-SIFT method has a worse classification performance for low FPRs than the Harris-(19×19) method. When a minimum FPR is desired, the Harris detector combined with the 19×19 basic region descriptor is the preferred method. However, when a higher FPR is allowed, the Harris-SIFT method performs best. Table 3 shows that, for a FPR of 7%, a TPR of 94% is achieved.

5.3 Aspect Angle Dependence

To verify the aspect angle dependence of the evaluated methods, we split the data set into three groups of estimated aspect angles, as follows:

1. Side views: ± 0 –20 deg
2. Angular views: ± 20 –60 deg, with respect to side view
3. Front/rear views: ± 60 –90 deg, with respect to side view

For every group of aspect angles, we evaluated the classification performance of all methods to see how they depend on the aspect angle of the targets. The results are shown in Fig. 7.

Figure 7 shows that the performance of the 19×19 descriptor is significantly better for side views than the performance of the SIFT and steerable filters' descriptors, whereas for angular views, this is opposite. For the front/rear views, the performance of the 19×19, SIFT, and steerable filters' descriptors are comparable. For all aspect angles, the invariant-moments descriptor performs worst, indicating that the information in this descriptor is insufficient to effectively

Table 3 True-positive ratios of max-class ROC for descriptors combined with Harris detector.

FPR	19×19 region	Steerable filters	Invariant moments	SIFT
0.015	0.72	0.80	0.45	0.79
0.030	0.85	0.87	0.47	0.85
0.050	0.92	0.88	0.48	0.92
0.070	0.93	0.92	0.51	0.94
0.100	0.00	0.92	0.57	0.00
0.250	0.00	0.00	0.58	0.00
0.800	0.00	0.00	0.71	0.00
1.000	0.00	0.00	0.72	0.00
AUC	0.0422	0.0385	0.0218	0.0395
FPR ≤ 0.05

Fig. 6 Classification performance of all evaluated region descriptors: (a) Results for maximum selection, giving information about the actual classification performance, and (b) results for similarity score, giving information about distinctive quality of the descriptors for the classes in our experiment.

discriminate between the classes. Because the reference set consists of only side views and front/rear views, the classification performance for the angular views gives an indication of the invariance for aspect angle changes. From Fig. 7(b), we conclude that the SIFT and steerable filters descriptors are most invariant to aspect angle changes of the targets in our data set. This agrees with the invariant properties of these descriptors compared to the 19×19 basic region descriptor.

6 Discussion

To benchmark the classification performance, we compared our classification results to the results in the earlier TNO paper.² We calculated a single point in the max-class ROC curve corresponding to a confusion matrix in the paper. For a FPR of 6%, the TPR is 93%. Although the comparison to only one point on the ROC curve is not very accurate, we used it as a comparison for the classification performance and plotted this together with our results in Fig. 8.

Compared to the earlier algorithm, we address the following differences:

1. Normalization of images: Previously, the same data set was used, but without normalization of the images. Compensation for differences in scale and intensity between images is left to be solved by the descriptors, and hence, depends on the invariance of the descriptors for these differences.
2. Location of key points: In our algorithm, the location of key points of the targets is part of the matching algorithm. It is not possible to match key points that have similar descriptors but have different locations on the targets.
3. Front and rear views: We expanded our reference set with front and rear views to make classification of all aspect angles possible.
4. Descriptor distance measure: We used the Euclidean distance as similarity measure for the descriptors,

Fig. 7 Classification performance for Harris detector, combined with different descriptors for different aspect angles of targets in data-set images.

Fig. 8 Comparison to performance of the earlier TNO algorithm, near the top right of the graph.

where previously the earth-movers (EM) distance was used. Using the EM distance could improve our classification results, because it makes matching of descriptors more robust for errors in key-point location.

For the normalization of data-set images, we chose a basic image size of 71×31 , based on the largest images in the data set. When we would have chosen a different image size, probably the optimal descriptor size would also have been different. When a different data set would be used, the basic file size should be reevaluated, based on the largest images in the data set. Downscaling of images means information is thrown away, which is undesirable.

We expected the steerable filter descriptor to perform better than the 19×19 descriptor because the steerable filters should enhance the edges in the region around each key point, leading to a better matching result. However, we have seen that this is not the case. An explanation could be that the steerable filter descriptor is constructed by concatenation of four filter responses. A strong response of one of the filters can be compensated by a weak response of one of the other filters. However, the choice to incorporate four filter responses is to make the descriptor more robust. In our case, edges are the important features; thus, the two anisotropic filters in the steerable filter descriptor could be left out to improve the performance of the descriptor. Furthermore, we used the standard filter size of 3 for the filters. The descriptor performance could be improved by optimization of the filter size because this depends on the resolution of the evaluated regions.

With SIFT, we have seen interesting results. The standard SIFT algorithm performed worst, together with the blob detector combined with the invariant-moments descriptor. However, the combination of the SIFT descriptor with the Harris detector showed best performance, together with the 19×19 descriptor. This shows that the invariant quality of the standard SIFT algorithm depends on the quality of the images. When used for small images, with little texture and low contrast, such as the images in our data set, the strongest characteristics are the edges. The DoG detector in the stan-

dard SIFT algorithm mainly detects blobs. Key points found on edges are eliminated. This explains the poor results for the standard SIFT algorithm in our evaluation.

Compared to classification of small sea targets in a real-world scenario, there are more than three classes of small targets and there is clutter, which makes classification more difficult. This makes our experiment not representative for a real operational scenario. Nevertheless, we have proven that classification of small targets using region descriptors gives satisfactory results for the defined data set. The algorithm should be tested with a larger data set, containing more types of small boats to see if it could be used in practice.

To use our algorithm for automatic classification of small targets onboard naval ships, an important aspect is the computation time for the classification of an infrared frame. It is not necessary to classify each frame of an infrared video stream, but classification of a frame should not be too slow. In the decision-making process to decide if a target is hostile, every second counts. To raise the class probability, several frames per second should be processed so that possible false alarms can be voted out. Using a dual core processor of 2.67 GHz and 4 Gb of memory, classification of one image from our data set takes 0.33 s using the Harris-(19×19) method. When using the Harris-SIFT method, classification takes 0.35 s and, for the Harris-steerable filters method, this is 1.3 s.

Although we showed that region descriptors perform quite well, other methods using shape characteristics may be preferred. Global methods, such as shape context, could be suitable for our data set. Generally, global features can be extracted faster, decreasing computation time.

7 Conclusions

In this paper, we have presented an experimental evaluation of region descriptors, when used for automatic classification of small sea targets in infrared video. The goal was to find out how the classification performance depends on the choice of key-point detector and region descriptor for the targets in our data set.

The performance of our algorithm is comparable to the earlier TNO algorithm.² However, we incorporated some techniques to make the algorithm more robust, as follows:

1. We used image normalization to make the algorithm robust to differences in scale and intensity of the targets.
2. The relative context of key-point location prevents matching of key points with dissimilar locations.
3. By adding front/rear views to the reference set, classification of all aspect angles is possible without reduction of the classification performance.

Furthermore, we conclude that the choice of key-point detector is crucial to get an acceptable classification result. Compared to the blob and SIFT key-point detector, only the Harris detector results in an acceptable classification performance. This indicates that the shape of the objects in the images is the most distinctive feature for classification. On the basis of the characteristics of the images in the data set and the time needed for classification, the preferred descriptors to use are the 19×19 basic region descriptor and the SIFT descriptor.

From the evaluated descriptors, the invariant-moments descriptor is least suitable for classification of the targets in our

data set. The information in the low-dimensional descriptor is insufficient to effectively discriminate between classes.

Acknowledgments

The experiment was supported by TNO's V602 program on Electro-Optical Sensor Systems.

References

1. P. B. W. Schwing, S. P. van den Broek, and M. van Iersel, "EO system concepts in the littoral," *Proc. SPIE* **6542**, 654230 (2007).
2. S. P. van den Broek, H. Bouma, M. A. C. Degache, and G. J. Burghouts, "Discrimination of classes of ships for aided recognition in a coastal environment," *Proc. SPIE* **7335**, 73350W (2009).
3. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(10), 1615–1630 (2005).
4. G. Carneiro and A. D. Jepson, "Phase-based local features," in *Proc. of 7th European Conf. on Computer Vision*, pp. 282–296 (2002).
5. H. Bouma, D. J. de Lange, S. P. van den Broek, R. A. W. Kemp, and P. B. W. Schwing, "Automatic detection of small surface targets with electro-optical sensors in a harbour environment," *Proc. SPIE* **7114**, 711402 (2008).
6. C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of Alvey Vision Conf.*, pp. 147–151 (1988).
7. P. Kovesi, "Harris corner detector," School of Computer Science & Software Engineering, University of Western Australia (2005).
8. D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **2**(60), 91–110 (2004).
9. D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. of 7th Int. Conf. on Computer Vision*, Kerkira, Greece, pp. 1150–1157 (1999).
10. M. K. Hu, "Visual pattern recognition by moment invariants," *Inf. Theory, IRE Trans.* **8**(2), 179–187 (1962).
11. W. Freeman and E. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(9), 891–906 (1991).
12. A. Vedaldi, "Extract SIFT features," Department of Computer Science, UCLA Vision Lab (2006).
13. P. Withagen, K. Schutte, A. M. Vossepoel, and M. Breuers, "Automatic classification of ships from infrared (FLIR) images," *Proc. SPIE* **3720**, 180–187 (1999).
14. J. A. Alves, "Recognition of ship types from an infrared image using moment invariants and neural networks," MS Thesis, NAVAL Postgraduate School, Monterey (2001).
15. S. Dudani, K. Breeding, and R. McGhee, "Aircraft identification by moment invariants," *IEEE Trans. Comput.* **26**(1), 39–46 (1977).
16. A. Reeves, R. Prokop, S. Andrews, and F. Kuhl, "Three-dimensional shape analysis using moments and Fourier descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.* **10**(6), 937–943 (1988).

Martijn M. Mouthaan joined the Royal Netherlands naval college in 1996 and graduated in 2001. After this, he went to the submarine service and sailed on a submarine as a junior technician officer for two years. In 2003, he joined the maintenance department on the Valkenburg Naval Air Station, where he was head of a maintenance crew and coordinator of the on-the-job training for new aircraft technicians. In 2006, he joined a team of project engineers at the Defence

Materiel Organisation in The Hague, responsible for the requirements management of the new Joint Support Ship for the Royal Netherlands Navy. During this period, he completed the master program of Media and Knowledge Engineering at Delft University of Technology. This paper describes the research he performed for his thesis project.

Sebastiaan P. van den Broek received his MSc and PhD in applied physics at Twente University in 1992 and 1997, respectively, doing research on modeling electrical and magnetic signals resulting from brain activity. He has been working in the Electro-Optics Group, the Netherlands Organization for Applied Scientific Research (TNO) since 1998. His interests in the field of image processing range from detection, tracking, and classification, extending to multisensor

fusion and sensor networks.

Emile A. Hendriks received his MSc and PhD, both in physics, from the University of Utrecht in 1983 and 1987, respectively, and he became an assistant professor with the EECMS faculty of the Delft University of Technology. In 1994, he became a staff member of the Information and Communication Theory Group (currently Pattern Recognition and Bioinformatics Group) of this faculty, and since 1997, he has headed the computer vision lab of this group as an associate professor. His interests are computer vision, low-level image processing, image segmentation, stereoscopic and 3-D imaging, motion and disparity estimation, structure from motion/disparity/silhouette, and real-time algorithms for computer vision applications.

Piet B. W. Schwing received his MSc in astronomy in 1983 and PhD in mathematical sciences in 1988, both from the University of Leiden, where from 1983 to 1987 he worked in the field of astrophysics. He studied dust properties and star formation in the Magellanic clouds using Infrared Astronomical Satellite (IRAS) data at IR wavelengths. Since 1987, he has been a scientist and project manager in the Electro-Optics Group, the Netherlands Organization for Applied Scientific Research (TNO). At present, he is program manager of Electro-Optical Sensor Systems and Improved Recognition. His interests are the detection of objects in IR surveillance, the characteristics and reduction of background clutter, and sensor system evaluation.