Synthetic template: effective tool for target classification and machine vision

Kaveh Heidary
Department of Electrical Engineering and Computer Science
Alabama A&M University
Normal, Alabama 35762, USA

***Abstract*—A process for replacing a voluminous image dictionary, which characterizes a certain target of interest in a constrained zone of effectiveness representing controlled states including scale and view angle, with a synthetic template has been developed. Synthetic template (ST) is a spatial map (grayscale image) obtained by combining the set of zone-specific training images that are ascribed to the target of interest. It has been shown that the solo-template ST correlation filter outperforms filter banks comprised of multiple target-class training images. A geometric interpretation of the basic ST concept is employed in order to further explain and substantiate its properties**.

Keywords—machine vision; image procession; target classification; correlation filter

I. Introduction

Machine vision involves the process of autonomous assessment of imagery data in a wide array of applications ranging from robotic navigation to biometrics and automatic detection and tracking of targets [1-5]. Computers are utilized to detect, identify and track objects of interest based on their electromagnetic or acoustic signatures, which may be expressed as two-dimensional data arrays obtained through an assortment of modalities including visible, infrared and synthetic aperture radar (SAR) imagery [6]. In the context of this paper a digital image is the result of spatial sampling and quantization of the filtered light energy emanating from the object and impinging on the focal plane of the optical sensor. Low level machine vision involves processing at the pixel level including image operations such as smoothing, enhancement and edge detection. Midlevel machine vision moves beyond pixels and involves larger abstractions such as shape, geometry and texture based classification, and high level vision involves context cognition including image understanding and interpretation [7-13]. This paper addresses a midlevel machine vision problem involving the development of a supervised learning algorithm for imagery based classification of objects. The goal is to develop a classifier that can determine the presence and location of the object of interest in arbitrary two-dimensional images. The classifier is constructed using a set of training images that represent the object of interest under assorted object states and viewing conditions that characterize the classifier's intended zone of effectiveness. The resultant classifier must be robust, in the sense of its ability to detect and locate with high reliability the object of interest under arbitrary view conditions within the intended zone of effectiveness. It must also be computationally efficient in terms of its ability to operate on large image files with low latency, using readily available hardware platforms. The two requirements of locating ability and computational efficiency of the desired classifier point in the direction of Fourier filtering [14].

The optimal procedure for determination of the presence of a known signal in the input waveform generated by the sensor, whose output is potentially corrupted by an additive stationary noise process, is matched filtering [15-17]. The matched filter is the optimal linear signal processor in the sense of maximizing signal-to-noise ratio (SNR) at the detector output. Under special circumstances, where the power spectral density of the noise process is uniform (white noise), the impulse response of the matched filter is equivalent to the time/space-reversed version of the sought after signal. The optimal method for detecting the presence of a known signal in the input waveform which is corrupted by an additive white Gaussian noise (AWGN) process is therefore cross-correlating the waveform with a replica of the signal of interest.

Pattern matching, where a window containing an image of the sought after object (target) is slid over the image under test, is a conventional approach to locating targets of interest in the input image [18-20]. Generally, a typical target of interest is characterized by many windows comprising the target-class training set of images. Each window pertains to the target image under a specific view condition such as scale, pose, lighting, possible partial obscuration, view and illumination directions, etc. An actual target in real images can render countless patterns due to variations in range (scale), both in-plane and out-of-plane rotation (pose), environmental conditions including lighting, shadow and partial obscuration effects [21-27]. Any given object can cast infinitely may different projections upon the sensor's focal plane array and therefore can produce countless images. Object image variability may arise from intrinsic and extrinsic inconsistencies. Intrinsic effects include object deformation, articulation and pose, and extrinsic effects include range, view angle, lighting and obscuration.

One of the elements of any machine vision system is a target dictionary of images associated with each object of interest. In a robust system, a typical target dictionary consists of numerous windows (target images), and the sensor image must be tested against all the windows in order to establish the target's presence and potential locations or lack thereof in the input image. The sensor images must be tested against

[*] Partial sponsorship for K.H. was provided by Department of Defense through US Army RDECOM W911NF-13-1-0136.

[*] E-mail: kaveh.heidarry@aamu.edu, Telephone: +1 256 372 5587.

.

numerous dictionaries, each containing huge numbers of images, where each dictionary represents a distinct target of interest. The size of the database constituting the image dictionaries for potential objects of interest in practical scenarios can be enormous and may involve tens of thousands of images [28]. Clearly, the memory and processing requirements placed on the system makes this approach impractical, especially for real-time applications. Processing sensor images in real-time with several target dictionaries each containing numerous images places an insurmountable barrier to practical autonomous vision systems.

In order to reduce the arduous computational burden of storing and processing vast image dictionaries, arising from the object image variability effects caused by intrinsic and extrinsic inconsistencies, synthetic discriminant functions (SDF) and distortion tolerant filters have been developed [29-46]. In the vein of SDF, this paper presents a novel and straightforward technique for substantially reducing the number of images in the target dictionary without adversely affecting robustness of the system. Reducing the number of images contained within the target dictionary results in proportionately smaller memory space dedicated to its storage and the abbreviated computational complexity. Implementation of the proposed algorithm can potentially lead to more economical machine vision systems with higher accuracy, lower storage and processing overhead and the concomitant reduced latency, smaller footprint, and lower power consumption.

As stated above, characterizing a certain target of interest under wide ranging target states and viewing conditions requires an inordinate number of training images (templates), and the associated immense storage and processing hardware. In practice, the viewing condition range of concern for a particular target of interest is partitioned into several tightly bound domains in the scale-rotation space. The universal target dictionary is comprised of a set of domain-specific dictionaries, each containing several target images. Henceforth in this paper, target dictionary refers to the domain-specific dictionary described above, whose elements are target renditions under constrained variations in scale and rotation (both in-plane and out of plane). It is noted that the target dictionary images, although very similar to each other due to their tightly bound domain origins, are nevertheless different from each other. In practical systems each domain-specific dictionary may indeed contain a single image due to the small number of available training images. In this paper, however, we assume that each domain-specific dictionary contains multiple images. We propose a method to distill all the training set images into a new virtual image and replace the multi-image target dictionary with the generated synthetic template (ST). In the operation phase, in order to establish the presence and location of the target of interest in the sensor image, rather than computing the cross correlations of the input image with respect to all the target dictionary images, it is correlated with the single-template ST. This results in storage and processing savings proportionate to the number of images in the original target dictionary.

The excellent performance of the basic synthetic template filter, in comparison to the bank of templates, suggests that formulating 3D models of the targets of interest and producing many computer generated images of each target spanning the respective scale and rotation ranges may be advisable in some applications. For example, in order to capture the full extent of image variability due to the relative positions and orientations of the target and senor as well as lighting conditions in a certain scenario, the 3D space of range, depression angle, and aspect angle is partitioned into the desired number of bins (zones of effectiveness). A large set of target images pertaining to each bin are generated using various scale-pose-view-lighting permutations, which are subsequently combined to construct the corresponding ST. The experimental results suggest that raising the number of model based images leads to performance enhancement without increasing the computational load of the classifier in the operation phase.

The remaining pats of this paper are organized as follows. The problem formulation is described in Section II. Test results pertaining to the performance of the basic ST classifier and comparisons to the full and partial banks of matched filters are presented in Section III. Section IV puts forward a geometric interpretation of the ST theory and presents illustrative simulation results. Concluding remarks and suggested future work are provided in Section V.

II. Problem Formulation

Let us assume the target dictionary contains several grayscale training images constituting the bank of templates (BT). The largest spatial dimensions of the training images along orthogonal directions are denoted as Δ_x, Δ_y. All the training images are zero padded in order to make their dimensions along the x-y axes equal to Δ_x and Δ_y, respectively. The training images are initially normalized with respect to their mean values, and each of the mean-compensated images are subsequently normalized with respect to the square root of integral of the square of the respective image intensity as shown in Eqs. 1-3.

$$\bar{s}_n(x,y) = s_n(x,y) - \bar{I}_n \;;\; 1 \le n \le N \tag{1}$$

$$\bar{I}_n = \oiint s_n(x,y)dxdy \tag{2}$$

$$\hat{s}_n(x,y) = \frac{\bar{s}_n(x,y)}{\sqrt{\oiint \bar{s}_n^2(x,y)dxdy}} \tag{3}$$

$$BT = \{\hat{s}_n(x,y) : 1 \le n \le N\} \tag{4}$$

Where, $s_n(x,y), \bar{I}_n$, $\bar{s}_n(x,y)$, $\hat{s}_n(x,y)$, denote, respectively, a typical zero-padded training image, mean value of the image, mean-compensated image, and the normalized image. Here and henceforth the surface integrations are with respect to the image surface. The number of images in the training set is N, and BT in Eq. 4 denotes the bank of templates.

The mutual cross correlation surfaces amongst all the normalized images of BT and the respective peak cross correlations are computed as follows.

$$\Lambda_{m,n}(x,y) = \oiint \hat{s}_m(x',y')\hat{s}_n\big((x'-x),(y'-y)\big)dx'dy' \;;$$
$$1 \le m,n \le N \tag{5}$$

$$\lambda_{m,n} = \max_{x,y}\left[\Lambda_{m,n}(x,y)\right] \tag{6}$$

where, $\Lambda_{m,n}(x,y), \lambda_{m,n}$ denote, respectively, the cross correlation surface and the corresponding peak cross correlation between two BT images $\hat{s}_m(x,y)$ and $\hat{s}_n(x,y)$. The function $\hat{s}_n(x,y)$ in Eq. 5 is periodically extended along both spatial directions.

One of the templates of BT which has the largest minimum peak cross correlation with respect to all the other images is declared the prototype template (PT) as shown below.

$$PT \in BT \tag{7}$$

$$\exists\ k: \ \min_n[\lambda_{k,n}] \geq \min_n[\lambda_{l,n}] \quad \forall l; \\ 1 \leq k, l, n \leq N \tag{8}$$

$$PT = \hat{s}_k(x,y) \tag{9}$$

Where, $\hat{s}_k(x,y)$ denotes PT. If there are multiple images that satisfy the condition of Eq. 8, one is chosen randomly and is declared as PT.

All the BT images are spatially aligned with respect to PT. This is done by spatially shifting each image such that the peak of its cross correlation surface with respect to PT occurs at (0,0).

$$\forall n \ \ \exists\, (x'_n, y'_n) \ : \ \ \Lambda_{k,n}(x'_n, y'_n) \geq \Lambda_{k,n}(x,y)\,; \\ \forall\, (0 \leq x \leq \Delta_x\,, 0 \leq \mathrm{y} \leq \Delta_y) \tag{10}$$

$$\widehat{w}_n = \hat{s}_n((x - x'_n), (y - y'_n)) \tag{11}$$

$$\widetilde{BT} = \{\widehat{w}_n(x,y) : 1 \leq n \leq N\} \tag{12}$$

Where, $\widehat{w}_n$ is a typical spatially shifted and renormalized target template, and $\widetilde{BT}$ denotes the bank of spatially shifted and aligned templates.

Summing all the spatially aligned target templates of Eq.12, and normalizing the resultant synthetic image such that the integral of its square is unity, one arrives at the synthetic template (ST).

$$h(x,y) = \textstyle\sum_{n=1}^{N} \widehat{w}_n(x,y) \tag{13}$$

$$\hat{h}(x,y) = \frac{h(x,y)}{\sqrt{\oiint h^2(x,y)dxdy}} = \\ = \frac{h(x,y)}{\sqrt{N + 2\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}\lambda_{m,n}}} \tag{14}$$

$$ST = \hat{h}(x,y) \tag{15}$$

Where, $\hat{h}(x,y)$ represents the ST.

As explained above, BT in Eq. 4 consists of the entire set of normalized training images, whereas PT in Eq. 9 contains only one of those images. PT is deemed the best representative of the training set in the sense that it has the largest minimum peak correlation with respect to all the trainers. A conceptual and geometric account of PT proceeds as follows: in the manifold of the training images within the image hyperspace, PT is closer to the center of gravity of the training manifold than any other BT image. Similar to PT, ST also consists of a single template, however, it is none of the actual training images, and is synthesized by amalgamation of all the trainers. A conceptual and geometric account of ST proceeds as follows: in the manifold of the training images within the image hyperspace, ST is indeed the center of gravity of the training manifold. ST is the center of mass of the convex hall of the training images. The spatial dimensions of the constituent templates of BT, PT, and ST are identical, namely $\Delta x \times \Delta y$. The computational complexity of processing a typical sensor image with BT is therefore higher than processing the same image with PT and ST, by a factor equal to the number of training images N. In the operation phase, the correlation of the filter template is computed with respect to the input image, and if the correlation value exceeds the user-specified threshold, the presence of target is declared at the specified location of the input image.

It is noted that BT is the conventional matched filter bank under the white noise assumption. In the next section the performance of correlation filters based on the full bank of templates (BT) is compared to those based on two single-image templates, namely the prototype template (PT) and the synthetic template (ST). Also, the performance of the correlation filter based on the fractional bank of templates (FBT) is examined. The images comprising a typical FBT are obtained by random selection of a user-prescribed number of images from BT.

III. Test Results

In this section the performance of correlation filters based on BT, PT, ST, and FBT are examined using actual images for training and testing. The images used in the test scenarios presented here are obtained from the Amsterdam Library of Object Images (ALOI), details of which are provided in [47] and the actual image databases are found at [48]. Many tests were conducted, where designated image sets pertaining to certain user-specified objects were utilized as the target-class training set of images. The images of the chosen target-class object which were not employed in the training process as well as the images of other non-target objects were utilized as the test set of images. Four types of correlation filters were constructed, as described in Section II, using the sequestered training set of images. The correlation filters were subsequently used as binary classifiers in order to classify each of the images in the test set as either target-class or non-target-class. The performance of each type of filter is characterized in terms of its receiver operating characteristic (ROC), where the probability of detection P_D is plotted in terms of the probability of false-alarm P_{FA}. In the tests presented here, P_D and P_{FA} for a particular classification filter refer to the proportion of the test images that are, respectively, correctly labeled and mislabeled by the corresponding filter.

Target-class and non-target-class training and test images employed in the first experiment were the image masks pertaining to objects number - 2, 550, 700, 800, and 950, which denote, respectively, *lab-keys, winny-the-pooh, daffy-duck, tea-can, and bananas* in ALOI. The database contains 71 image masks for each object, all taken at the same range and at equally spaced view angles in $[0 - 350^{\circ}]$. In this experiment, the target-class universe of images (TCUI) comprises 24 object-2 images with equally spaced view angles in $[0-115^{\circ}]$, and the non-target-class universe of images (NTCUI) comprises 284 images equally distributed between ALOI objects- 550, 700, 800 and 950. The reason for restricting TCUI to a prescribed subset of the object-2 image set is to ensure that the target-class manifold in the image hyperspace is a simply connected zone. The view angles $[0-115^{\circ}]$ and uniform scale in this example constitute the classifier zone of effectiveness. For applications where the domains of view angle and scale cover wider ranges, the TCUI manifold may have to be partitioned into multiple simply connected zones in the hyperspace, and a particular classifiers must be devised for each simply connected zone. The focus of this paper, however, is the design of a binary classifier whose zone of effectiveness in the hyperspace of images is a simply connected region.

Figure 1 shows samples of the image sets associated with each of the five objects used in this experiment, where each row contains five views of the same object. Minimum peak correlation among the 24 images of TCUI is 0.555, and maximum peak correlation between TCUI members on one hand and NTCUI on the other hand is 0.75917. Ten images of TCUI are randomly selected in order to create the training set of images, form which four types of correlation filters are constructed. Each type of filter is then employed to classify members of the 298-image test set comprised of the remaining 14 target-class and all 284-non-target class images. It is noted that the training was based solely on ten target-class images and none of the test images were involved in the training process. The simulation was repeated 200 times, each time randomly selecting a ten-image subset of TCUI, constructing four types of binary classifiers, namely BT, PT, ST, FBT, and utilizing each classifier in order to label 298 non-trained-on test images. The resultant P_D and P_{FA} parameters for each classifier were then averaged across 200 trials. Figure 2 shows one instantiation of the training set of images comprised of ten randomly selected images from TCUI, and the computed ST for a typical simulation run.

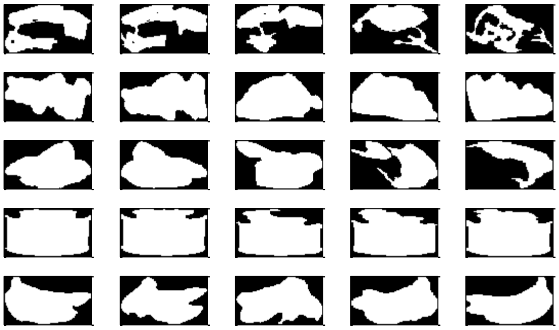

Fig. 1. Samples of target-class objects are shown in the top row. TCUI is limited to 24 equally spaced view angles in $[0-115^{\circ}]$ of object-2. Rows two through five show samples of non-target-class objects pertaining to object types 550, 700, 800, 950, respectively. The training set is formed by random selection of ten images from TCUI, and the test set is formed by the remaining 14 target-class and all 284 non-target-class images.

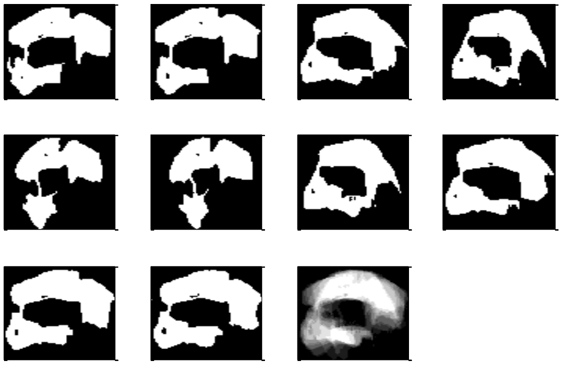

Fig. 2. The target-class training set of images consisting of ten object-2 images are shown in the top two rows and the two left figures in the bottom row. These images, which pertain to a typical simulation run, are randomly selected from TCUI and comprise BT. The right image in the bottom row is the corresponding ST.

The performance of each binary classifier is characterized in terms of its respective ROC. For a typical filter, setting the threshold level at 1 results in P_D=P_{FA}=0, and as the threshold level is lowered both P_D and P_{FA} increase. In the experiments presented here, for each filter the threshold was lowered until P_D=1 was achieved. The classifier performance results are plotted in Figs. 3 and 4. The plots of Fig. 3 show that ST performance is far superior than PT. These plots also shows that for low P_{FA} values, ST outperforms BT in terms of achieving higher P_D for the same P_{FA} value, even though its computational complexity is lower by a factor of ten. The performance of ST was also compared to those of different FBTs and the results are plotted in Fig. 4. As explained before, the images constituting each FBT are obtained by randomly selecting a user-prescribed number of images from the ten-image BT. For each test case, multiple permutations were conducted by randomly selecting the prescribed number of training images, constructing the FBT, computing the ROC of the resultant classifier and averaging the results across multiple permutations. Plots of Fig. 4 show the performance comparisons between ST and three different FBTs comprised of one, five, and eight training images. It is seen that ST outperforms the one and five-image FBTs by great margins. It is also seen that for low P_{FA} values ST is superior to the eight-image FBT. Comparing the PT performance result shown in Fig. 3 with that of the one-image FBT (M=1) in Fig. 4, it is seen that PT is clearly superior. This result is supported by intuition, because PT and FBT with M=1, although both consist of one target-class template each, PT has a distinct property that makes it a better filter. PT is chosen in order to minimize its distance to the center of gravity of the training manifold, whereas the FBT (M=1) is chosen randomly. Table 1 lists the probabilities of detection and false-alarm for different classifiers. It is seen that ST performs better than FBT with seven images by yielding higher P_D and lower P_{FA} concurrently. It is also seen that, for very low values of P_{FA}, ST outperform even the full bank of templates (BT), whose computational complexity is ten times that of ST.

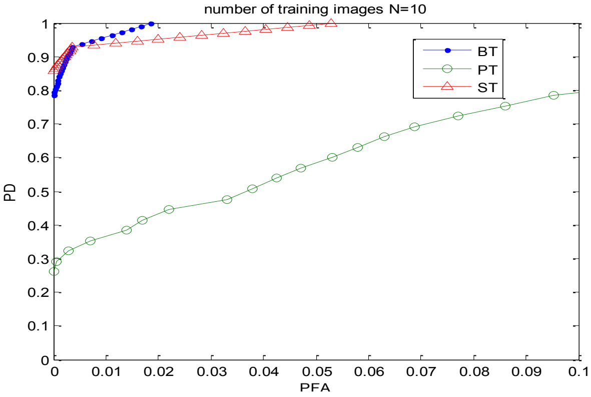

Fig. 3. Receiver operating characteristic (ROC) plots for BT, PT and ST. The training set of images is comprised of N=10 target-class (ALOI object-2) images. Test set of images consists of 14 target-class and 284 non-target-class (objects-550,700,800,950) images. The top-left corner represents perfect recognition and the diagonal line (not shown) connecting (0,0) to (1,1) denotes chance.

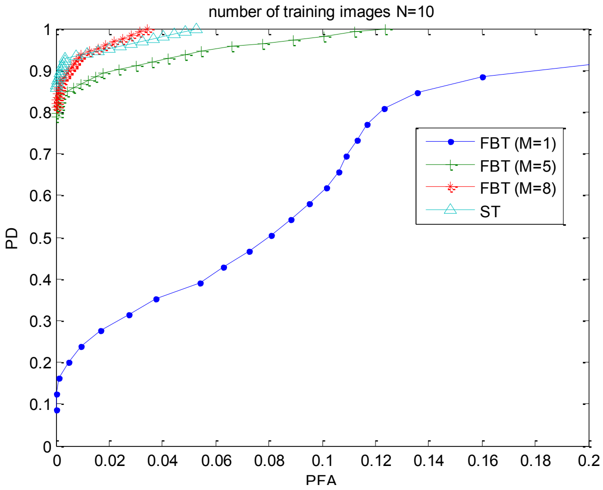

Fig. 4. Receiver operating characteristic (ROC) plots for ST and three FBTs with one, five and eight templates. The training set of images is comprised of N=10 target-class (ALOI object-2) images. Test set of images consists of 14 target-class and 284 non-target-class (objects-550,700,800,950) images.

TABLE I. List of several P_D-P_{FA} pairs for classifiers based on the synthetic template (ST), prototype template (PT), bank of templates (full bank BT), and three FBTs.

	ST	PT	FBT (M=1)	FBT (M=4)	FBT (M=7)	BT
P_D	0.8571	0.2609	0.087	0.6	0.7647	0.7857
P_{FA}	0	0	0	0	0	0
P_D	0.875	0.3533	0.8098	0.8	0.8039	0.8393
P_{FA}	0	0.007	0.1233	0.003	0.0003	0.0011
P_D	0.9048	0.5072	0.8859	0.8833	0.8922	0.9018
P_{FA}	0.0024	0.038	0.1604	0.0247	0.0033	0.0027
P_D	0.9524	0.9384	0.9620	0.95	0.951	0.9554
P_{FA}	0.02	0.1879	0.3098	0.0812	0.021	0.0092

The next test scenario involves target-class and non-target-class training and test images pertaining to object masks 9, 23, 33, 58, and 75 which denote, respectively, *shoe, blue-bear, chess-horse, blue-car, and boat* in the ALOI database. Figure 5 shows five sample images of each object. Twenty-seven object-9 images corresponding to equally spaced view angles in $[0-130^\circ]$ constitute TCUI, and 284 images of the other four objects constitute NTCUI. There are 71 images for each non-target-class object corresponding to equally spaced view angles in $[0-350^\circ]$. The zone of effectiveness of the classifier in this example includes view angles $0-130^\circ$ at the same range. The minimum peak correlation among 27 images of TCUI is 0.7739, and the maximum peak correlation between TCUI on one hand and NTCUI on the other hand is 0.8697. The training set of images is formed by randomly selecting ten target-class images from the 27-image TCUI. As before, the training process does not utilize any non-target-class images. The 17 remaining target-class and 284 non-target-class images comprise the test set of images. As explained earlier, utilizing the training set of images four types of binary classifiers are constructed. Each classifier is then employed to label 301 previously unseen test images. The simulation was repeated 200 times, where each run involved forming new training and test sets of images as outlined above, constructing four types of binary classifiers, and labeling the test images. The performance of each type of classifier was characterized by averaging the respective ROCs across 200 simulation runs. Figure 6 shows ten object-9 images comprising a single instantiation of the training set of images involved in a particular simulation run, and the respective ST. In the simulation run of Figure 6, BT is comprised of the entire set of ten trainers shown, PT is one of the trainers whose minimum peak correlation with respect to the remaining nine is maximum, and FBT consists of M<10 randomly selected images from the set of ten trainers. Each filter is then utilized to classify each image in the test set as target or non-target. Plots of Fig. 7 show the performance characteristics of the binary classifiers. It is clearly seen that the performance of ST classifier is superior to all the other filters. Remarkably the ST-based filter performs even better than the full-set bank of templates. In this example, utilization of ST results in a classifier which requires ninety-percent less storage and ninety-percent less processing compared to BT, yet it is more robust.

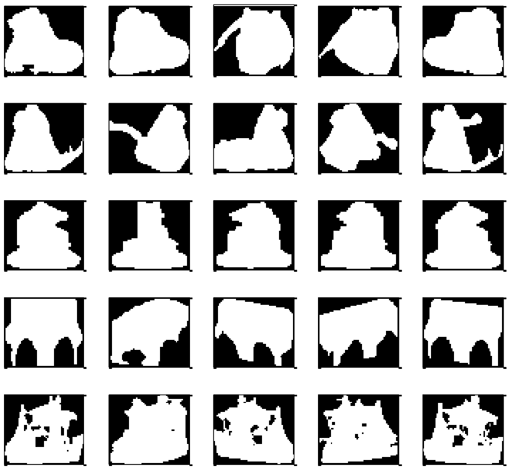

Fig. 5. Samples of target-class images are shown in the top row. TCUI is comprised of 27 images of the type-9 object. Rows two through five show

samples of non-target-class images pertaining to object types 23, 33, 58, 75, respectively.

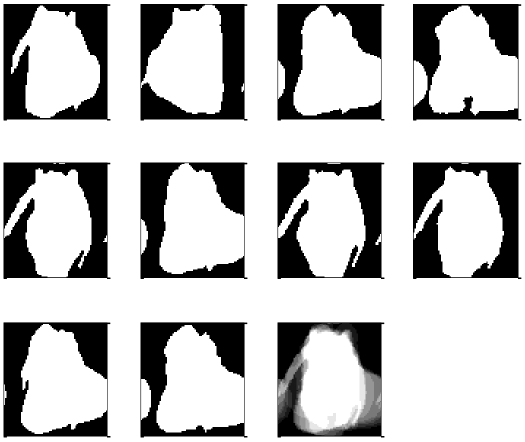

Fig. 6. Images in the top two rows and the left two columns of the bottom row constitute the training set of images. These ten object-9 images are randomly selected from the 27-image TCUI and comprise one instantiation of BT. The right image of the bottom row shows the corresponding ST.

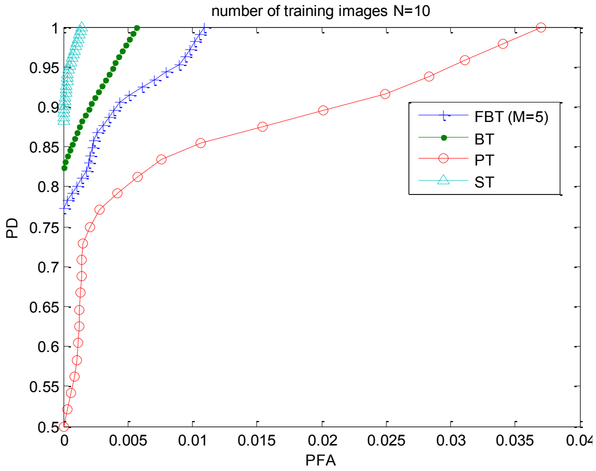

Fig. 7. Receiver operating characteristic (ROC) plots for FBT with (M=5), BT, PT and ST. The training set of images is comprised of N=10 target-class (ALOI object-9) images. Test set of images consists of 17 target-class and 284 non-target-class (objects-23,33,58,75) images. The top-left corner represents perfect recognition and the diagonal line (not shown) connecting (0,0) to (1,1) denotes chance.

The test results presented in this section are typical of numerous performance assessment findings that were obtained in conjunction with various experimental campaigns conducted using object images derived from the ALOI database under various scenarios. In every case, the solo-template ST performed better than the single-template PT, as expected. The classifier performance metrics include the overall ROC assessments as well as comparison of the probabilities of correct classification, for various classifier filters, at fixed user-prescribed values of probabilities of misclassification. The ST also outperformed FBTs with multiple templates, which is somewhat counter intuitive. What is most remarkable, however, is the fact that for low P_{FA} values ST consistently outperforms the full bank of matched filters (BT) upon which it is based. This means that combining multiple target templates and forming the synthetic template, results in concurrently higher computational efficiency, measured in terms of lower memory and reduced complexity, and superior classifier performance. In order to explain what initially appears to be a counter intuitive phenomenon, we embark on a geometric interpretation of the synthetic template concept, which is presented in the next section.

IV. Geometric Interpretation

This section provides a simple geometric interpretation of the basic ST theory, where images are represented as vectors in a hyperspace. The training set of images is a manifold in the image hyperspace, and the classifier comprises a set of hyper-spheres with equal radii. A plausible explanation of the impressive performance of the basic ST classifier, observed under various assessment scenarios and test results partially presented in Section III, is given using a simplified 2D vector analogy.

A. Image-Point Analogy

Let us assume that all images of potential interest inhabit a hyperspace, where each image is uniquely represented by a point. Let us also assume, for the purpose of explanatory simulations presented here, that the 2D analogy to the above image hyperspace is the xy-plane of the Cartesian coordinate system, and each (x,y) point is the 2D version of a unique image. The distance between two points in the plane has an inverse relationship to the peak normalized correlation between the two respective images.

$$D_{p_i,p_j} = \alpha \frac{1-\lambda_{I_i,I_j}}{\lambda_{I_i,I_j}} \tag{16}$$

Where, p_i and I_i denote, respectively, a point in the xy-plane and the corresponding image in the hyperspace, D_{p_i,p_j} is the Euclidean distance between two points in the plane, λ_{I_i,I_j} is the peak value of the normalized cross correlation surface between two corresponding images, and α is the user prescribed proportionality constant.

$$\lambda_{I_i,I_j} = \max_{u,v}\left[\frac{\iint I_i(u',v')I_j\big((u'-u),(v'-v)\big)du\prime dv\prime}{\sqrt{\iint I_i^2(u,v)dudv \iint I_j^2(u,v)dudv}}\right] \tag{17}$$

Where, (u,v) represents a point in the plane of a particular image $I_i(u,v)$, which itself is represented as a single point in the image hyperspace, and its 2D counterpart is (x,y).

In machine learning, a set of images representing an object class, say *chair*, are used in order to create a binary classifier (filter) which is capable of distinguishing, in new images, *chair* from other objects *(non-chair)*. one of the standard techniques to accomplish this task is template matching. A target dictionary comprised of a large number of training images of *chair* is produced, and a threshold level is set. In the test phase, the class label of the image under test is determined in accordance to its peak cross correlation values with respect to the templates contained within the prearranged target dictionary. If the peak cross correlation of the test image with respect to any one of the target dictionary images exceeds the threshold, the test image is classified *chair (target)*, otherwise it is classified *non-chair*. In practice, the design process places stringent limits on the type of *chair* as well as the sensor view conditions such as range, elevation and azimuth angles, and lighting, in assembling the template dictionary. The cluster of training images which constitute the target dictionary, is a meticulous subset of the entire set of known target images. Target dictionary elements are all assumed to have high mutual peak correlations. In the *chair* case, for example, if the large set of known target class images represent many different types of chair with different view angles and scales, the set is judiciously partitioned into multiple clusters (zones of effectiveness), each containing several highly correlated images. It is this cluster of target (*chair)* images with high mutual peak correlations that constitutes our training set from which the binary target filter (classifier) is derived. The filter, therefore, has a very limited zone of effectiveness. The universal binary classifier, capable of recognizing *chair*, is comprised of a larger number of such filters, each with a limited zone of effectiveness also called the target zone. This discussion is concerned with a cluster of tightly bound target class training images that we call bank of templates (BT). Conceptually, BT is represented by a set of points cloistered inside a small volume in the image hyperspace. The trainer manifold in the image hyperspace is assumed to be a simply connected domain. The volume encompassing BT is inside a hyper-sphere.

$$R \leq \frac{1-\lambda_{min}}{\lambda_{min}} \tag{18}$$

where, λ_{min} denotes the minimum value of peak correlations $(i.e. \lambda_{min} = 0.85)$ among all the BT members. The target zone has a very small volume and an amorphous shape, and is contained within the volume of the sphere of Eq. 18. It is assumed that all points within the target zone (zone of effectiveness) represent *chair* and no other object. Let us assume, BT contains a large number of images, say N=20 or so. The filter comprising the fully populated bank of templates BT is a zonal classifier and is an element of the universal *chair* classifier. BT can be utilized for recognizing *chair* in its designated zone of effectiveness. In principle, the fully populated BT may be employed to recognize prospective manifestations of *chair*. This approach, however, may not be practical, due to the fact that the universal *chair* classifier can potentially contain many zonal classifiers, each with its own BT. The storage and processing requirements of using the universal *chair* classifier, comprised of many fully populated BTs, make this approach prohibitively expensive. Therefore, alternatives to the fully populated BT for the zonal classifier are sought.

One solution to the storage and processing problems caused by the large number of images contained within BT is to replace it with a single template that best represents BT, called the prototype template PT. A logical choice for PT, is to choose the template whose minimum peak correlation with respect to all the remaining BT templates is maximum. Geometrically, PT is the template that is closest to the center of the hyper-sphere of Eq. 18. This is an intuitively sensible solution, since all of the BT elements are very similar to each other, and choosing the one which has, on average, the greatest similarity to the group, as a whole, seems to be a rational choice. Replacing BT with PT reduces both the storage and processing requirements by factors of N, where N denotes the number of elements of BT. Another logical solution would be to amalgamate all the BT elements and form a synthetic template ST. In practice, this merging process is carried out by first properly scaling and spatially shifting all of the BT elements and then adding and rescaling the resultant image. It is noted that the image comprising ST is not a physical image. Similar to PT, replacing BT with ST reduces both the storage and processing requirements by factors of N. The third solution is to select a subset of the BT templates and form a fractional bank of templates FBT. Replacing BT with FBT reduces both the storage and processing requirements by factors of N/M, where N, M denote the numbers of templates in BT, FBT, respectively, and $M \leq N$.

The four zonal *chair* classifiers (filters), described above, are each comprised of one or multiple spheres in the image hyperspace. The BT, FBT, and PT filters are comprised, respectively, of N, M, and one spheres, each centered at the respective template. All spheres associated with a particular filter have equal radii. The ST filter is a single sphere, centered at a point which may not coincide with any of the actual templates in BT. The filter volume is the volume in the hyperspace that is contained within the volume(s) of the hyper-sphere(s) constituting the classifier. The hyper-sphere radii (thresholds) are chosen in order to strike the desired balance between probabilities of detection P_D and false-alarm P_{FA}. In this discussion, P_D denotes the proportion of the target zone volume that is contained inside the *chair* filter. P_{FA}, on the other hand, represents the non-target zone hyperspace volume that falls inside the *chair* filter. In order to make P_D=1, the radii of the spheres constituting the filter (BT, PT, ST, FBT) must be increased such that the entire target zone is contained within the filter volume, which may lead to unacceptably large P_{FA}. On the other hand, in order to make P_{FA}=0 one must decrease the radii until the non-target zone volume contained within the filter volume is vanished, which may lead to unacceptably small P_D.

In applications where data storage and processing speed are at a premium and one is forced to use a single template for the zonal filter, a choice between PT, ST, and FBT with M=1 has to be made. Contrary to intuitive considerations that the performance of PT and ST filters are comparable, we have found this to be a false assumption. In many test cases using real and simulated images, we have found that ST consistently outperforms PT by great margins. In every test we have conducted, it has been shown that ST has concurrently higher P_D and lower P_{FA} than PT. Considering that ST is obtained by merging all the N images in BT, it represents in the hyperspace a point which is closer to the center of gravity of the convex hull representing the target zone (zone of effectiveness) than any of the actual templates in BT. However, as the number of images in BT increases (say N=100), one would expect that PT and ST would have comparable P_D, P_{FA} performance, which turns out not to be the case. The synthetic template (ST) outperforms the prototype template (PT) even when the number of templates in BT is very large.

What is even more striking is the fact that in all the test cases that we have conducted, ST outperforms the fractional bank of templates (FBT) with substantial number of templates. In many test cases using actual images, we have shown that the one-template ST has superior performance, in terms of higher P_D and lower P_{FA}, than FBTs with ten or higher templates. This is indeed a remarkable feat, since by creating a synthetic template one can achieve a classification system with smaller memory requirement, lower latency and higher accuracy at the same time.

B. Simulation Results

In order to illustrate the detection capability of the synthetic template, in the following examples we use the xy-plane as the 2D representation of the image hyperspace. For ease of presentation we assume that the target zone is represented by the unit square area with corners at (0,0), (1,0), (0,1), and (1,1). Each point of the unit square represents a potential target image, and all the exterior points represent non-target images. It is assumed that N points in the unit square are labeled as target and constitute the known target set of images. It is noted that the training process is oblivious to the fact that the target zone is comprised of the unit square. Rather, all it knows is: the N points it is given belong to the target class. In this example, therefore, the set of N labeled points represents the training set of images or the bank of templates BT. The fractional bank of templates FBT is obtained by randomly choosing a subset of BT. The prototype template PT is one of the BT elements, and is obtained as the point whose maximum distance with respect to all other BT points is minimum. The synthesized prototype ST, on the other hand, is a point in the unit square whose coordinates are means of the respective coordinates of the BT members. It is noted that ST, in general, does not coincide with any of the BT points. The binary classifiers based on BT, FBT, PT and ST are each comprised of one or multiple disks in the xy-plane. Each disk is centered at a respective template (actual or synthesized point), and all the disks comprising a certain classifier have identical radii (thresholds).

An unlabeled test point (image) is classified as target if it is inside any of the disks comprising the classifier, otherwise it is labeled as non-target. In order to assess the performance of a certain classifier the areas of the unit square and the non-target area (outside the unit square) that fall inside the classifier's constituent disks are computed. Probabilities of detection and false alarm, P_D and P_{FA}, are equal to, respectively, the areas of the unit square and the outside-region that are contained within the disks. Figure 8 illustrates the 2D analogy of the image hyperspace, the target zone manifold and the binary classifier.

The performance of four types of classifiers described above were studied by conducting the following simulation. A user-prescribed number of points (i.e. N=100) were randomly selected from the unit-square target area. The set of N labeled points forms the training set. The bank of templates BT consists of N disks centered at these target points with equal radii r.

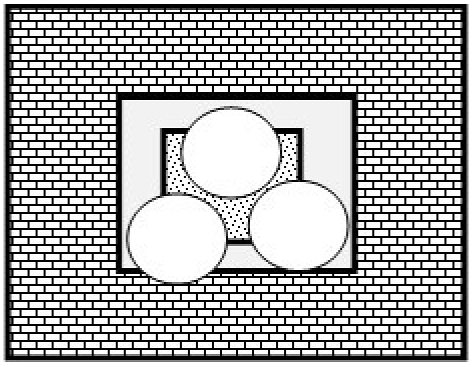

Fig. 8. The interior dotted square region represents the target zone, the gray annular region around the target zone is the exclusion zone, where no images can exist, and the exterior brick region represents the non-target universe. The circles constitute a binary classifier comprised of a bank of three templates, obtained from a potentially larger set of known target images. The target and non-target areas overlapping the circles represent, respectively, P_D and P_{FA}.

Clearly, setting r=0 results in $P_D=P_{FA}=0$. Increasing r will result in raising P_D, while P_{FA} remains zero as long as none of the disks protrude from the unit square. In virtually all cases $P_{FA}=0$ is possible only if $P_D<1$ can be tolerated. Likewise, $P_D=1$ is achieved at the expense of $P_{FA}>0$. Computing P_D and P_{FA} for various values of r and plotting the result, one obtains the receiver operating characteristic (ROC) of the classifier. This is done by repeating the simulation many times, randomly selecting N training points each time, computing the respective P_D and P_{FA} pairs for various r values, and averaging the results across all trials.

In each simulation round, a subset of the BT's N training points consisting of M<N points are randomly chosen to form the FBT. One of the N training points which has the smallest maximum distance with respect to the remaining N-1 points, is chosen to form PT. A new point is synthesized by computing the means of the respective coordinates of the N points comprising BT to form ST. As before, target filters consist of one or multiple disks with equal radii and centered at the corresponding points. Similar to the BT classifier, the ROC plots for FBT, PT, and ST classifiers are computed.

In the example of Fig. 9 the number of training points was set at N=25 and the simulation was repeated for 100 trials. In each trial run, BT consists of 25 randomly selected points in the unit square area constituting the target zone, and PT and ST are derived from the corresponding BT. The performance of each filter is computed by averaging the ROC results across 100 trial runs for the respective classifier. It is seen that the one-template ST clearly outperforms the 25-template BT, which is somewhat consistent with the results we have obtained using actual images in Section III. In this simulation, however, PT performs better than BT for all P_{FA} values, which is contrary to the experimental results of Section III. The reason for this apparent paradox is the fact that, in the simulations of Fig. 9, the non-target region abuts the target region. This implies that potential non-targets may have peak correlations with respect to potential targets, that approach $\lambda=1$. In practice, however, this is not the case.

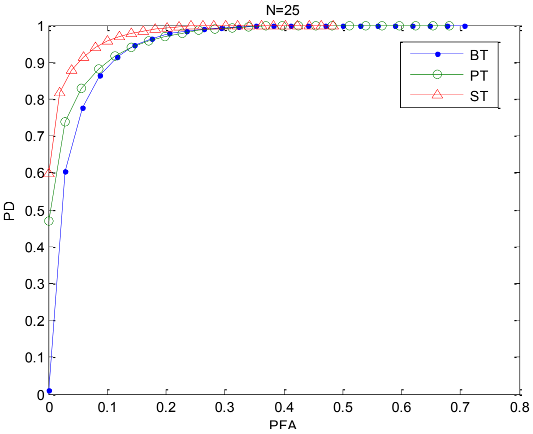

Fig. 9. Performance comparisons for BT, PT, and ST. The number of training points is N=25, and there is no exclusion zone between target and non-target zones EZ=0.

In order to present a more realistic scenario, where none of the potential non-target elements have extremely high peak correlations with respect to target elements, a zone of exclusivity was established around the unit target area in the 2D example. An annular region with width of 0.2 around the unit square is assumed to be void of any target or non-target elements. All of the tests conducted with this scenario show that the single-template ST is superior to target dictionaries containing many templates. In the example of Fig. 10, BT consists of 25 points randomly selected from the target area. As expected BT has superior performance compared to that of PT.

Contrary to expectation, however, ST which consists of a single point, determined as the mean of BT points, outperforms BT. Performance of FBTs comprised of 5, 10 and 15 points randomly selected from the 25 BT points are also plotted. Fig. 11 shows the performance of various classifiers when number of trainers is set at N=10. The simulation results of Figs. 10 and 11 are consistent with the experimental results of Section III.

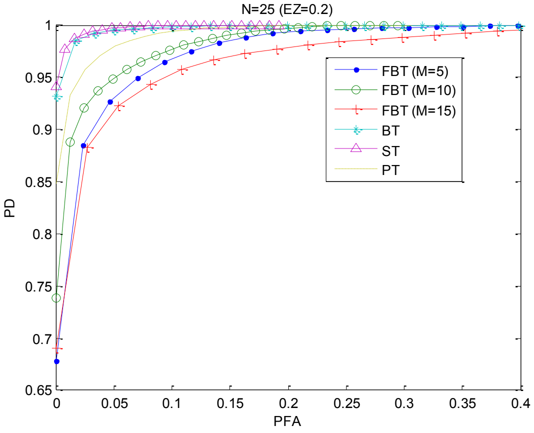

Fig. 10. Performance comparisons between BT, PT, ST, on one hand and FBTs with different number of templates on the other. Total number of known target points is N=25, and exclusion zone had a width of EZ=0.2.

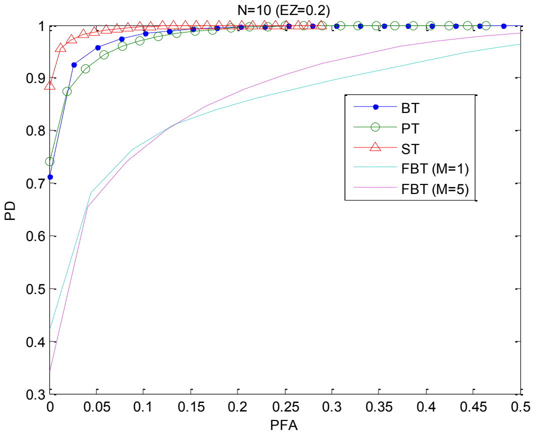

Fig. 11. Performance comparisons between BT, PT, ST, on one hand and FBTs with different number of templates on the other. Total number of known target points is N=10, and width of the exclusion zone is 0.2.

V. Conclusions

This paper paves the way towards developing a conceptually simple and computationally efficient mechanism for replacing voluminous target image dictionaries with much smaller sets of synthetic templates for target detection, classification and machine vision applications. Synthetic template (ST) is a spatial map (grayscale image) obtained by combining a set of training images that are ascribed to a target of interest. The rudimentary ST presented here is obtained by pixel-wise summation of the uniformly weighted, spatially shifted and normalized target-class training set of images. It constitutes a correlation filter that is used to determine the presence and locations of the target of interest in new images, or determine if a new image is that of the target of interest. It has been shown, using numerous test scenarios, that the solo-template ST outperforms filter banks comprised of multiple target-class training images. The ST classifier produces higher probability of correct classification and lower probability of misclassification than a large bank of target-class images (matched filters). The basic ST is generated offline in a straightforward manner and its online utilization results in lower system overhead in terms of abbreviated memory space requirement and reduced computational complexity, potentially leading to systems with more condensed physical footprint, lower power consumption, and reduced latency. Experiment based quantitative studies using many test scenarios with real images were carried out to assess the efficacy of ST and a representative sampling of the performance results are presented. An intuitive geometric interpretation of the basic ST theory and the corresponding simulation results provide a plausible explanation for its remarkable performance. In this paper, all target images within a particular dictionary are assumed to be highly correlated, and the dictionary is distilled into a single ST. In practice, where the target-class training set of images represent versatile and unconstrained views of the target, multiple dictionaries have to be created by suitable partitioning of the training set. Each appropriately created dictionary is then distilled into a single ST. Work on developing efficient algorithms for clustering the training set of images and automatic formation of target dictionaries is ongoing.

REFERENCES

[1] B. G. Batchelor (Editor), Machine Vision Handbook, Springer, 2012.

[2] C. Steger, M. Ulrich, C. Wiedemann, Machine Vision Algorithms and Applications, Wiley -VCH, 2007.

[3] E.R. Davies, Computer and Machine Vision: Theory, Algorithms, Practicalities, Academic Press, 2012.

[4] R. C. Gonzalez, R. E. Woods, Digital Image Processing, Prentice Hall, 2007.

[5] T. Acharya, A. K. Ray, Image Processing - Principles and Applications, Wiley, 2006.

[6] B. Bhanu, 'Automatic target recognition: State of the art survey,' IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-22, No. 4, (1986) pp. 364-379.

[7] H. Kauppinen, T. Seppanen, M. Pietikainen, 'An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification,' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 2, (1995) pp. 201-207.

[8] S. Belongie, J. Malik, J. Puzicha, 'Shape matching and object recognition using shape contexts,' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 24, (2002) pp. 509-522.

[9] Y. Rui, T.S. Huang, S.F. Chang, 'Image Retrieval: Current Techniques, Promising Directions, and Open Issues,' Journal of Visual Communication and Image Representation, Vol. 10, No. 1, (1999) pp. 39-62.

[10] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, H.J. Zhang, 'Image classification for content-based indexing,' IEEE Transactions on Image Processing, Vol. 10, No. 1, (2001) pp. 117-130.

[11] D. Zhang, G. Lu, 'Shape-based image retrieval using generic Fourier descriptor,' Signal Processing: Image Communication, Vol. 17, No. 10, (2002) pp. 825-848.

[12] [12] M. Bober, 'MPEG-7 visual shape descriptors,' IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 6, (2001) 716–719.

[13] Y. Liu, D. Zhang, G. Lu, W.Y. Ma, 'A survey of content-based image retrieval with high-level semantics,' Pattern Recognition, Vol. 40, No. 1, (2007) 262-282.

[14] A.B. Vander Lugt, 'Signal detection by complex filtering,' IEEE Transactions on Information Theory, Vol. 10, No. 2, (1964) pp. 139-145.

[15] G.L. Turin, 'An introduction to matched filters,' IRE Transactions on Information Theory, Vol. 6, No. 3, (1960) 311-329.

[16] A, Papoulis, Signal Analysis, McGraw-Hill, 1977.

[17] H. L. Van Trees, Detection, Estimation, and Modulation Theory, Vol. 1, John Wiley & Sons, 2001.

[18] Q.S. Chen, M. Defrise, F. Deconinck, 'Symmetric phase-only matched filtering of Fourier-Mellin transforms for image registration and recognition,' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 12, (1994) pp. 1156-1168.

[19] L.M. Novak, G.J. Owirka, C.M. Netishen, 'Radar Target Identification using Spatial Matched Filters,' Pattern Recognition, Vol. 27, No. 4, (1994) pp. 607-617.

[20] A. Mahalanobis, A.V. Forman Jr., N. Day, M. Bower, R. Cherry, 'Multi-class SAR using shift-invariant correlation filters,' Pattern Recognition, Vol. 27, No. 4, (1994) pp. 619-626.

[21] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, 'From few to many: Illumination cone models for face recognition under variable lighting and pose,' IEEE Transactions on Pattern Analysis and Machine Vision, Vol. 23, No. 6, (2001) pp. 643-660.

[22] K.J. Dana, B.V. Ginneken, S.K. Nayar, and J.J. Koenderink, 'Reflectance and texture of real world surfaces,' ACM Transactions on Graphics, Vol. 18, No. 1, (1999) pp. 1-34.

[23] P. N. Belhumeur, D. J. Kriegman, 'What is the set of images of an object under all possible illumination conditions?,' International Journal of Computer Vision , Vol. 28, No. 3, (1998) pp. 1-16.

[24] P. Hallinan, A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions, PhD thesis, Harvard University,1995.

[25] R.C. Hoover, A.A. Maciejewski, R.G. Roberts, 'Fast Eigenspace Decomposition of Images of Objects with Variation in Illumination and Pose,' IEEE Transactions on Systems, Man, and Cybernetics, Vol. 41, No. 2, (2011) pp. 318-329.

[26] R. Garg, D. Hao, S.M. Seitz, N. Snavely, 'The Dimensionality of Scene Appearance,' Proceedings IEEE 12th International Conference on Computer Vision (2009) pp. 1917-1924.

[27] T.M. Caelli, Z.Q. Liu, 'On the minimum number of templates required for shift, rotation and size invariant pattern recognition,' Pattern Recognition, Vol. 21, No. 3, (1988) 205-216.

[28] S. Landeau, T. Dagobert, 'Image database generation using image metric constraints: an application within the CALADIOM project,' Proc. SPIE 623410, (2006) pp. 1-12.

[29] D. Casasent, D. Psaltis, 'Position, Rotation, and Scale Invariant Optical Correlation,' Applied Optics, Vol. 15, No. 7, (1976) pp. 1795-1799.

[30] Y. N. Hsu, H.H. Arsenault, G. April, 'Rotation-invariant digital pattern recognition using circular harmonic expansion,' Applied Optics, Vol. 21, No. 22, (1982) pp. 4012-4015.

[31] D. Mendlovic, E. Marom, N. Konforti, 'Shift and scale invariant pattern recognition using Mellin radial harmonics,' Optics Communication, Vol. 67, No. 3, (1988) pp. 172-176.

[32] Y. Sheng, J. Duvernoy, 'Circular-Fourier-radial-Mellin transform descriptors for pattern recognition,' JOSA A, Vol. 3, No. 6, (1986) pp. 885-888.

[33] H. J. Caulfield, M. H. Weinberg, 'Computer recognition of 2-D patterns using generalized matched filters,' Applied Optics, Vol. 21, No. 9, (1982) pp. 1699-1704.

[34] B. V. K. Kumar, 'Minimum-variance synthetic discriminant functions,' JOSA A, Vol. 3, No. 10, (1986) pp. 1579-1584.

[35] A. Mahalanobis, B.V.K. Kumar, D. Casasent, 'Minimum average correlation energy filters,' Applied Optics, Vol. 26, No. 17, (1987) pp. 3633-3640.

[36] D. Casasent, G. Ravichandran , S. Bollapragada, 'Gaussian-minimum average correlation energy filters,' Applied Optics, Vol. 30, No. 35, (1991) pp. 5176-5181.

[37] D. Casasent, G. Ravichandran , 'Advanced distortion-invariant minimum average correlation energy (MACE) filters,' Applied Optics, Vol. 31, No. 8, (1992) pp. 1109-1116.

[38] M. Savvides, B.V.K.V. Kumar, 'Quad Phase Minimum Average Correlation Energy Filters for Reduced Memory Illumination Tolerant Face Authentication,' Lecture Notes in Computer Science, Vol. 2688, (2003) pp. 1056-1065.

[39] K. Heidary, H.J. Caulfield, 'Application of supergeneralized matched filters to target classification', Applied Optics Vol. 44 No. 1, (2005) pp. 47-54.

[40] K. Heidary, H.J. Caulfield, 'Nonlinear Fourier correlation,' Proceedings of SPIE Vol. 7340A, (2009) pp. 1-11.

[41] R. B. Johnson, K. Heidary, 'A unified approach for database analysis and application to ATR performance metrics,' Proceedings of SPIE, Vol. 7696Z, (2011) pp. 1-20.

[42] A. Mahalanobis, R.R. Muise, S.R. Stanfill, A. Van Nevel, 'Design and application of quadratic correlation filters for target detection,' IEEE Transactions on Aerospace and Electronic Systems, Vol. 40, No. 3, (2004) pp. 837-850.

[43] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, M. Yi, 'Robust Face Recognition via Sparse Representation,' IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 2, (2009) pp. 210-227.

[44] [44] K. Heidary, 'Distortion tolerant correlation filter design,' Applied Optics, Vol. 52, No. 12, (2013) pp. 2570-2576.

[45] E. Perez, B. Javidi, 'Nonlinear distortion-tolerant filters for detection of road signs in background noise,' IEEE Transactions on Vehicular Technology, Vol. 51, No. 3, (2002) pp.567-576.

[46] O. Arandjelovic, R. Cipolla, 'A methodology for rapid illumination-invariant face recognition using image processing filters,' Computer Vision and Image Understanding, Volume 113, No. 2, (2009) 159-171.

[47] J. M. Geusebroek, G. J. Burghouts, A. W. M. Smeulders, The Amsterdam library of object images, Int. J. Comput. Vision, Vol. 61, No. 1, (2005) pp. 103-112.

[48] http://staff.science.uva.nl/~aloi/ .