On Classifying Silhouettes in Adverse Conditions

Conrad Sanderson [1, 2] and Danny Gibbins [1, 2]
[1] *Electrical and Electronic Engineering, University of Adelaide, SA 5005, Australia*
[2] *CRC for Sensor Signal and Information Processing, Mawson Lakes, SA 5095, Australia*
conradsand@ieee.org danny@eleceng.adelaide.edu.au

Abstract

We compare the performance of holistic and local feature approaches for the purpose of classifying silhouettes in adverse conditions (i.e. occlusions by other silhouettes, noise and imperfect localization by a Region of Interest algorithm, resulting in clipping and scale changes). Holistic feature extractors based on Hu's moment invariants and Principal Component Analysis (PCA) are coupled with a classifier based on gaussian densities, while a local feature extractor based on the 2D Hadamard Transform (HT) is coupled with a Gaussian Mixture Model (GMM) based classifier. Experiments show that the HT/GMM approach is relatively robust to clipping, scale changes and occlusions; however, in its current form it is highly sensitive to noise. The results further show that the moment based approach achieves relatively poor performance in advantageous conditions and is easily affected by clipping and occlusions; the PCA based approach is highly affected by scale changes and clipping, while being relatively robust to occlusions and noise.

1. INTRODUCTION

Classification of objects based on their silhouettes is particularly useful when using Infra-Red (IR) imagery. While IR images are more suitable for operation in various environmental conditions than visible spectrum images [21], [29], there is still quite a lot of variability in the greyscale representation of objects; these variabilities can stem from the the object's history (e.g. due to variable use of engines), sea and air temperature (partly dependent on the time of day), location and range [21], [24]. To discount the effects of greyscale variations, binary versions of IR images can be used [13]. When only one solid object is present in a binary image, a *silhouette* of the object is visible; see Fig. 1 for an example.

In this paper we consider part of the problem of localizing a specific object. In particular, we assume that we are given a Region of Interest (ROI), on which we have to perform a two-class recognition task: either the presented ROI contains the object that we want (i.e. a *true object*), or it doesn't (i.e. an *impostor object*). The recognition task described above is known as a *verification* task (also known as a *detection* or an *authentication* task), and is in contrast to the *closed set identification* task, where a given object is assigned into one of K object classes (here K is the number of *known* objects). The closed set identification task represents a *controlled* environment and as such is not representative of our problem; the verification task represents an *uncontrolled* environment, where *any* type of object could be encountered [12].

The verification task can be present in two configurations of an object localization system:

1) Several regions are automatically found in a given image by a ROI locator algorithm; each ROI is assigned a verification score which represents the likelihood of containing the required object; the ROI with the highest score (which also exceeds a threshold) is then deemed to contain the required object.
2) Similar to configuration (1), but instead of using a ROI algorithm, the given image is split into many overlapped windows of various sizes; each window is considered to be a ROI [3].

When utilizing the first approach, the overall performance of the system can highly depend on the quality of the ROI algorithm [23]; besides the algorithm providing improperly located regions (e.g. objects at incorrect scale, partially shifted out view), it may *miss* the required object altogether. This is in contrast to the second approach, where the verification system subsumes the functionality of the ROI locator. The first approach can be considered to be a computationally simplified version of the second approach and hence may be preferable when computing hardware is limited.

A pattern recognition system can be generally split into two distinct parts: feature extraction and classification. Approaches to feature extraction from images containing objects can be roughly divided into three areas: boundary descriptions, local features (i.e. parts based descriptions) and holistic features. Holistic extractors utilize information from the entire image, while local feature extractors divide the image into several parts (possibly non-uniformly sized parts) and extract features for each part.

Examples of boundary descriptions include distances of each boundary pixel from the centroid [26] and Fourier Descriptors [10]; examples of parts based approaches include decomposition into distinct surfaces [19], regional Principal Component Analysis (PCA) [20], [23], and 2D Discrete Cosine Transform (DCT) [10], [25]; examples of holistic descriptions include moment invariants [8], [10], [14], [31], and holistic PCA [16], [27].

As noted before, local features describe only a part of the image of an object; hence when only a few parts of the object are corrupted we would expect only a small subset of the feature vectors to be affected. A possible disadvantage of local features is that relatively complex classifiers may be required in order to take advantage of the spatial relation between object parts (e.g. 2D Hidden Markov Models used in face classification [2]). Holistic representations, on the other hand, represent the entire image of an object using only one feature vector; the classifier can thus be relatively simple, e.g. an object's features can be considered to have a gaussian distribution. Further examples and discussion of object descriptors can be found in [10], [17], [32].

Several studies [5], [16], [24] have compared the performance of various approaches for recognizing objects in IR images. However, all of these studies assumed good quality ROI localization or used a closed set identification setup; furthermore, in [16], [24] silhouette representations of objects was not used. Hence the aim of this paper is to evaluate, on a common database, several combinations of feature extractors

Fig. 1: Left to right: size normalized object; corresponding silhouette; corrupted by speckle noise; clipped; scaled; overlapped with another silhouette.

and classifiers for the purposes of silhouette verification, subject to imperfect ROI localization and other adverse conditions.

We consider two holistic feature extractors, based on moment invariants and PCA, as well as a local feature extractor based on the 2D Hadamard Transform (HT) [10]. Two Bayesian classifiers are used: the first classifier assumes that the features for each class have a gaussian distribution, while the second assumes a distribution based on a mixture of gaussians. The performance of several feature/classifier combinations is evaluated under several conditions of the silhouette images: clean (i.e. no impediments), corrupted by speckle noise, partially shifted out of view (clipped), partially occluded by another object, and finally corrupted by a scale change. The last three corruptions represent an imperfect ROI locator, while the speckle noise corruption is a representative of imaging artefacts, a boundary corruption process (other than occlusions by another object), as well as a process which adds *previously unseen elements* to the image (i.e. elements of a different nature than what is present in training images).

The rest of this paper is organized as follows. In Section 2 we overview the three feature extraction techniques, followed by a brief description of the two classifiers in Section 3. Section 4 is devoted to experiments on classifying silhouettes in the abovementioned adverse conditions. The paper is concluded and future work is suggested in Section 5.

2. FEATURE EXTRACTION

In the following sections it is assumed that the binary image is described by $\{f(x,y) \in \{0,1\} \mid x = 0,1,\cdots,N_{\mathtt{X}}\text{-}1,\ y = 0,1,\cdots,N_{\mathtt{Y}}\text{-}1\}$, where $N_{\mathtt{X}}$ is the number of columns and $N_{\mathtt{Y}}$ the number of rows.

A. Hu's Moment Invariants

In this feature extraction method, *Hu's moment invariants* are derived from normalized central moments of a given image; it is often stated that these moments are invariant to translations, scale changes and in-plane rotations [10]. A moment of $(p+q)$-th order is defined as:

$$m_{pq} = \sum_{x=0}^{N_{\mathtt{X}}-1} \sum_{y=0}^{N_{\mathtt{Y}}-1} x^p y^q f(x,y) \tag{1}$$

for $p,q = 0,1,2,\cdots$. *Central moments* are in turn defined as:

$$c_{pq} = \sum_{x=0}^{N_{\mathtt{X}}-1} \sum_{y=0}^{N_{\mathtt{Y}}-1} (x-\mu_x)^p (y-\mu_y)^q f(x,y) \tag{2}$$

where $\mu_x = m_{10}/m_{00}$ and $\mu_y = m_{01}/m_{00}$. *Normalized central* moments are in turn defined as: $n_{pq} = c_{pq}/c_{00}^{\gamma}$, where $\gamma = (p+q)/2 + 1$ and $(p+q) \geq 2$. Central moments address translations of an object, while normalized central moments in turn address the scale of the object (as c_{00} describes the number of white pixels in a binary image). Seven moment invariants (ϕ_1 to ϕ_7) can then be obtained as non-linear combinations of second and third order normalized central moments [8], [10]; the combinations address the problem of in-plane rotations. As the values of the resultant features can be very small, it is customary to utilize the logarithm of their absolute values in order to avoid precision problems [14]. Formally, a feature vector based on the seven features is formed using:

$$\mathbf{x}^T = [\ \log|\phi_i|\ ]_{i=1}^{7} \tag{3}$$

There are two main problems with this feature extraction method. First, if a part of the object becomes invisible (e.g. due to partially shifting out of view), or there is some noise present, or part of another object is visible, the values of μ_x, μ_y and c_{00} are affected, leading to ϕ_1 to ϕ_7 being affected; secondly, the features are correlated, due to the basis functions defined by the monomials defined in (1) and (2) not being orthogonal [15].

B. Principal Component Analysis

In Principal Component Analysis (PCA) based feature extraction [7], [30], the first step is to concatenate all the columns of the binary image to form a high dimensional vector; let us denote the resultant vector as $\mathbf{r}$. A new feature vector, usually with a lower dimensionality, is then obtained using:

$$\mathbf{x} = \mathbf{U}^T (\mathbf{r} - \mathbf{r}_\mu) \tag{4}$$

where $\mathbf{U}$ contains D eigenvectors (corresponding to the D largest eigenvalues) of the training data covariance matrix, and $\mathbf{r}_\mu$ is the mean of training vectors. D has the following constraints: $D \leq N_T$, where N_T is the number of training vectors, and $D \leq N_{\mathtt{X}} N_{\mathtt{Y}}$. If $D = N_{\mathtt{X}} N_{\mathtt{Y}}$ then no dimensionality reduction occurs; in that case, vector $\mathbf{x}$ represents a decorrelated version of $\mathbf{r}$. A method for choosing the dimensionality is given in Sec. 4-D.

As this feature extraction technique basically produces dimensionality reduced versions of binary images, we would expect it to be affected by scale changes, clipping and rotations; the onus of any robustness to these changes would be taken by the classifier.

C. 2D Hadamard Transform

As opposed to the holistic feature extraction methods presented in Sections 2-A and 2-B, where analyzing *one* image results in *one* feature vector, in the 2D Hadamard Transform (HT) based approach we obtain a *set* of feature vectors from one image. In the literature this type of feature extraction is known as a *local feature* approach and as a *parts based* approach [2], [18].

The 2D HT is similar in nature to the 2D DCT [10] (which is the heart of the JPEG compression algorithm [28]); we shall interpret the HT as a "black and white" version of the DCT, and postulate that it is a good candidate for processing binary images.

In a similar manner to DCT based feature extraction [25], a given image is analyzed on a block-by-block basis; each block overlaps neighbouring blocks by a configurable amount of pixels. Each image block $\alpha(k,l)$, where $k,l = 0,1,\cdots,N_{\mathtt{P}}-1$ and $N_{\mathtt{P}} = 2^n$, is decomposed in terms of orthogonal 2D Hadamard basis functions (see Fig. 2). The result is an $N_{\mathtt{P}} \times N_{\mathtt{P}}$ matrix $H(u,v)$ containing 2D Hadamard coefficients:

$$H(u,v) = \frac{1}{N_{\mathtt{P}}} \sum_{k=0}^{N_{\mathtt{P}}-1} \sum_{l=0}^{N_{\mathtt{P}}-1} \alpha(k,l)\ (-1)^{\beta(k,l,u,v)} \tag{5}$$

where

$$\beta(k,l,u,v) = \sum_{i=0}^{n-1} [\ b_i(k)p_i(u)\ +\ b_i(l)p_i(v)\ ] \tag{6}$$

$$b_i(k) = \ i\text{-th bit in the binary representation of } k \tag{7}$$

$$p_i(k) = \begin{cases} b_{n-i}(k) & \text{for } i = 0 \\ b_{n-i}(k) + b_{n-(i+1)}(k) & \text{for } i \in [1, n-1] \end{cases} \tag{8}$$

The summations in Eqns. (6) and (8) are performed in modulo 2 arithmetic. As the 2D HT is similar in nature to the

2D DCT, the basis functions can be thought of as representing different "frequency" components. If we assume low frequency information to be dominant, we can order the resulting coefficients according to a zig-zag pattern [10], reflecting the amount of information stored in each coefficient (see Fig. 3); formally, for a block located at (k, l), the Hadamard feature vector is composed of:

$$\mathbf{x}^{(k,l)} = \left[\begin{array}{cccc} h_0^{(k,l)} & h_1^{(k,l)} & \ldots & h_{M-1}^{(k,l)} \end{array} \right]^T \tag{9}$$

where $h_n^{(k,l)}$ denotes the n-th Hadamard coefficient and M is the number of retained coefficients. By truncating the number of coefficients we are effectively throwing out high frequency information. The vectors from the entire image can then be collected in a set:

$$X = \{ \mathbf{x}_i \}_{i=1}^{N_V} \tag{10}$$

where the superscript indicating the location of each source block has been replaced by the subscript i, which indicates the i-th vector of the image.

3. CLASSIFIERS

A. Gaussian Based

In the gaussian based classifier each class is assumed to have a gaussian distribution, i.e.:

$$P(\mathbf{x}|\lambda) \triangleq \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{\exp\left[\frac{-1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \tag{11}$$

where $\lambda = \{ \mu, \mathbf{\Sigma} \}$ is the parameter set, D is the dimensionality, $\mathbf{\Sigma}$ is the covariance matrix and μ is the mean.

Let us assume we have two parameter sets λ_C and λ_I, which describe the distributions of object C's vectors and impostor objects' vectors respectively. A given object is classified as object C when $\Lambda(\mathbf{x}|\lambda_C, \lambda_I) \geq t$ or as an impostor object when $\Lambda(\mathbf{x}|\lambda_C, \lambda_I) < t$. Here t is a tunable decision threshold and $\Lambda(\mathbf{x}|\lambda_C, \lambda_I)$ is a log-likelihood ratio:

$$\Lambda(\mathbf{x}|\lambda_C, \lambda_I) = \log P(\mathbf{x}|\lambda_C) - \log P(\mathbf{x}|\lambda_I) \tag{12}$$

Note that in (12) we assumed non-informative prior probabilities of the two classes.

Assuming there is enough data, the classifier is trained by taking the mean and covariance matrix for a class to be the sample mean and sample covariance matrix of the training vectors for that class, respectively. However, in our case there is only a few training images for each object class (see Sec. 4-A), resulting in only a few training vectors when using holistic feature extraction; this in turn makes the estimation of covariance matrices for each object ill advised. The following alternative training strategy is used. First, we define a *world model* as a model representing any object and find its mean vector and covariance matrix using vectors representing all training true objects; due to the relatively small amount of training data, the covariance matrix is restricted to be diagonal. The mean for a particular object is taken to be the mean of that object's vectors; the covariance matrix for each object is then inherited from the world model [7]. We postulate that the world model is a good representation of many objects and use it to represent impostor objects.

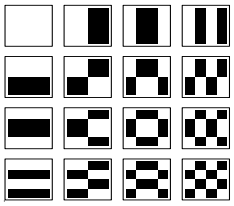

Fig. 2: 2D Hadamard basis functions for N_P=4; white areas indicate $+1$ while black areas indicate -1.

u \ v	0	1	2	3
0	0	1	5	6
1	2	4	7	12
2	3	8	11	13
3	9	10	14	15

Fig. 3: Zig-zag ordering of 2D Hadamard coefficients $H(u, v)$ for N_P=4.

B. Gaussian Mixture Model Based

The Gaussian Mixture Model (GMM) based classifier [7], [22], [30] can model distributions much more precisely compared to the single gaussian classifier, as each class is now assumed to have a distribution comprised of a mixture of gaussians:

$$P(\mathbf{x}|\lambda) \triangleq \sum_{g=1}^{N_G} w_g \mathcal{N}(\mathbf{x}|\mu_g, \mathbf{\Sigma}_g) \tag{13}$$

Here $\lambda = \{ w_g, \mu_g, \mathbf{\Sigma}_g \}_{g=1}^{N_G}$, N_G is the number of gaussians and w_g is the weight for gaussian g (with $\sum_{g=1}^{N_G} w_g = 1$ and $\forall\, g : w_g \geq 0$). The higher the N_G, the more precise the model (assuming enough training data); moreover, even if diagonal covariance matrices are utilized, it is possible to model correlated data as long as $N_G \geq 2$ [22].

In a similar manner to the gaussian classifier described in Sec. 3-A, training is accomplished as a two-stage process. First, a world model is estimated using the Expectation Maximization (EM) algorithm [4]; training data from all true objects is utilized. A model for each class is then obtained by *adapting* the world model using a form of Maximum *a Posteriori* adaptation [9], [22]. As for the gaussian classifier case, we postulate that the world model is a good representation of many objects and thus use it to represent impostor objects.

As stated in Sec. 3-A, there are only a few training vectors available when using holistic feature extraction; as this isn't enough data for estimating more than one gaussian, the GMM approach is only useful when dealing with 2D HT based feature extraction, where it is possible to obtain a large set of feature vectors from each image.

For mathematical convenience we shall assume that each vector in the set $X = \{ \mathbf{x}_i \}_{i=1}^{N_V}$ is independent and identically distributed (iid), leading to the following re-definition of the log-likelihood ratio [c.f. (12)]:

$$\Lambda(X|\lambda_C, \lambda_I) = \sum_{i=1}^{N_V} \log P(\mathbf{x}_i|\lambda_C) - \sum_{i=1}^{N_V} \log P(\mathbf{x}_i|\lambda_I) \tag{14}$$

Note that when using 2D HT based feature extraction, the above definition implies that the spatial relation between all blocks is lost. We thus expect some robustness to translations of the object, as well as minor deformations. Moreover, if the object is partially occluded (or otherwise corrupted), only some of the blocks will be affected; as long as the contribution of $\log P(\mathbf{x}|\lambda)$ for affected blocks does not adversely affect the sums, the HT/GMM approach should also be robust to occlusions.

4. EVALUATION

A. Database and Associated Experiment Protocols

The database is comprised of 64 synthetically generated objects (32 of which are ships); each object is rotated (out-of-plane rotation) 360^o in 5^o steps, resulting in 72 images per object. Each image has a resolution of 512×512 pixels and contains the object against a black background; the object is represented as greyscale pixels.

The views are split into three disjoint sections: train, evaluation and test. Taking the canonical side-on view to be the 0^o view, images for 0^o and $+5^o$ are assigned to the train section, while images for -5^o were assigned to the evaluation section. In the experiments reported in this paper, images for $\pm 10^o$ are used as test images (see Table 1). It must be noted that in this particular configuration we have assumed that images for 0^o, $\pm 5^o$ and $\pm 10^o$ are various realizations of the "side-on" view.

The objects are also split into three disjoint sections: *true objects*, *evaluation impostor objects* and *test impostor objects*. 48 objects are assigned to the true objects section,

while 16 objects are assigned to evaluation impostors and test impostors sections (eight to each section).

For all experiments, the train section was utilized as a source of data for training the object models; the evaluation section was used for tuning feature extractors and classifiers. Once the optimum parameters are found, the test section is used for final performance measurement.

This database partitioning is necessary to avoid optimistic biases in the performance evaluation; moreover, the evaluation section is necessary in order to avoid overfitting (i.e. to ensure good generalization capability) [7], [30].

B. Performance Measures

There are two types of errors that can occur in an object verification system: a *false acceptance* (FA), which occurs when the system accepts an *impostor object*, or a *false rejection* (FR), which occurs when the system refuses a *true object*. The performance of verification systems is generally measured in terms of *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR), defined as:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of impostor object presentations}} \quad (15)$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of true object presentations}} \quad (16)$$

To aid the interpretation of performance, the two error measures are often combined using the Half Total Error Rate (HTER), defined as HTER=(FAR+FRR)/2; the HTER is a special case of the Decision Cost Function [1], [6]. A particular case of the HTER, known as the Equal Error Rate (EER), occurs when the system is adjusted (e.g. via tuning a threshold) so that FAR=FRR on a particular data set.

C. Pre-Processing

Prior to feature extraction or applying any image corruption, all objects are normalized in size as follows. We find the mean (μ) and standard deviation (σ) of object pixel positions along the x and y dimensions of the image; the object is then assumed to be contained within the area specified by $\mu \pm 2.5\sigma$ in each dimension; 2.5 was chosen so that all objects in the database fit into the specified area; the area is extracted and normalized to have a size of 128 columns and 64 rows. Thresholding is used to generate a binary representation. An example of a size normalized object and the resulting silhouette is shown in Fig. 1.

The size normalization stage corresponds to a size normalized image being provided by either the ROI stage or the exhaustive window scanning approach.

D. Experiments

The amount of training and evaluation data was artificially increased by using mirrored versions of images. In each experiment the classifier was given the model of the object we are interested in and test images of that object and impostor objects; each given image was classified as either containing the object we want (i.e. a true object), or containing a different object (i.e. an impostor object). This procedure was repeated for all objects in the *true objects* section, resulting in a total of 48×2=96 true object presentations and $48\times8\times2$=760 impostor object presentations.

In the first experiment we evaluated the performance on clean data of all combinations of holistic features with the gaussian based classifier. The GMM based classifier was used only with 2D HT features; based on preliminary experiments, a block size of 16×16 was chosen with a block location advance of two pixels (resulting in an 87.5% overlap). The relatively small block advance was used to ensure an adequate amount of training data is available (as the larger the overlap, the more feature vectors are obtained from an image); this procedure may also have a beneficial side-effect: the components of each object are in effect "sampled" at various degrees of translations, resulting in models which should be robust to translations of the objects.

TABLE 1: DATABASE CONFIGURATION.

angle	true objects (48)	evaluation impostor objects (8)	test impostor objects (8)
0^o, $+5^o$	training data	-	-
-5^o	evaluation data	evaluation data	-
-10^o, $+10^o$	test data	-	test data

Preliminary experiments with the GMM based classifier showed that using all extracted blocks lead to poor performance. The problem was traced back to the EM training algorithm for the world model, where blocks which were the most similar to each other tended to be modeled by a small number of gaussians with a relatively high weight. While normally this kind of modeling is desirable, in our case the most similar blocks tended to be of the black background and the solid areas of the object; these areas carry no discriminant information. The database also contains objects which have elongated sections (e.g. a ship), resulting in a set of blocks which are the same, which in turn also caused the EM algorithm to put an emphasis on modeling one particular section of the object.

To reduce the effects of blocks with the same content, a post-processing heuristic was added: for each object's set of feature vectors, any repetitious feature vectors are removed; moreover, any feature vector whose elements are all zero is removed (i.e. vectors representing the black background). This post-processing method causes the number of vectors extracted from an image to be dependent on the content of the image; hence the ratios resulting from Eqn. (14) are no longer comparable across different objects. To address this problem, the ratio in (14) was redefined to:

$$\Lambda(X|\lambda_C,\lambda_I) = \frac{1}{N_V}\sum_{i=1}^{N_V}\log P(\mathbf{x}_i|\lambda_C) - \frac{1}{N_V}\sum_{i=1}^{N_V}\log P(\mathbf{x}_i|\lambda_I) \quad (17)$$

In other words, the effect of variable number of feature vectors is discounted by using average log-likelihoods.

The evaluation section of the database was used to *jointly* optimize the performance of the feature extractors and classifiers (e.g. dimensionality of feature vectors and thresholds); the optimization was performed in terms of minimum EER on evaluation data.

For PCA based feature extraction, the dimensionality was varied from 1 to 128, doubling the number of dimensions in each step (i.e. 1, 2, 4, $\cdots$, 128). The optimum dimensionality, according to performance on evaluation data, was found to be 64 (i.e. the dimensionality of binary images was reduced from 8192 to 64, representing a reduction of 99.2%).

For HT based feature extraction, each possible dimensionality was based on the cumulative amount of coefficients along the diagonals traced by the zig-zag pattern (see Fig. 3). The number of gaussians in the GMM approach was varied from 1 to 128, doubling the number of gaussians in each step (i.e. 1, 2, 4, $\cdots$, 128); the optimum number of gaussians always turned out to be less than 128. The optimum dimensionality for 16×16 blocks was 15 and the corresponding number of gaussians was 64.

Once the optimum parameters were found on the evaluation section, the resulting systems were tested on clean images (i.e. non-corrupted) from the test section. Results are presented in Table 2. Out of the two holistic feature extractors, moment invariants obtain the worst performance; as the moments are correlated, their poor performance can be partly attributed to the diagonal covariance matrix assumption used in the gaussian classifier. The HT/GMM combination achieved performance comparable to that of the gaussian classifier utilizing PCA derived features.

The results for images corrupted by speckle noise, clipping, scale changes and occlusions are shown in Figs. 4 to 7. For the speckle noise results, the noise level indicates the percentage of pixels which were randomly set to either zero or one; the location of the pixel to be corrupted was randomly selected according to a uniform distribution. For clipping experiments, the shift level indicates the fraction of columns by which the object has been shifted to the right. For scale experiments, the scale level indicates the size multiplier of each object. For occlusion experiments, the overlap level indicates the fraction of columns which have been corrupted by a secondary silhouette, moving in from the left; the secondary silhouette was taken to be one of the test impostor objects.

The results show that all three approaches are affected by noise, with the HT/GMM approach the most sensitive. The sensitivity of moment invariants stems from the fact that as more noise is present, the more affected μ_x, μ_y and c_{00} are (see Sec. 2-A).

Analysis shows that the high sensitivity of the HT/GMM approach is partially due to the post-processing heuristic, which utilizes only rudimentary rules for determining whether a given feature vector is discriminative. For noisy images, vectors from blocks containing noise (e.g. only one white pixel) are passed to the classifier, which in turn contribute to the averages in Eqn. (17). Moreover, due to the nature of the block-by-block analysis, vectors from blocks which contain noise-like patterns (e.g. a block containing one white pixel in the bottom right corner) are used during the modeling stage. This causes the object models to represent noise-like patterns, which in turn means that vectors from noisy blocks tend to give relatively high likelihoods, from both the client and world models. The differences in the averages in Eqn. (17) are thus reduced, leading to worse performance.

The results further show that the PCA/gaus. approach is very sensitive to translations and scale changes. Hu's moment invariants are somewhat less affected, due to their explicit translation and scale normalization steps; however, as soon as part of each object is moved out of view, the performance of moment invariants rapidly decays (the reasoning for this degradation is given in Sec. 2-A). This is in contrast to the HT/GMM approach where the object can be moved by more than 40% of the image width before the performance is considerably affected. The HT/GMM approach is also affected by scale changes, but considerably less than the holistic approaches. The relative robustness of the HT/GMM approach to translations can be attributed to the use of the parts based representation and the loss of spatial relation between each part (i.e. the exact location of each part of an object has little or no influence). We conjecture that the relative robustness to scale changes stems from some object parts changing relatively little. An example of this would be the mast of a ship joining the ship's hull, represented by the shape "⊥"; if we split this shape into two parts, we obtain "⌟" and "⌞"; while scale changes would in effect move the two parts, their geometrical nature would stay constant.

TABLE 2: PERFORMANCE USING NON-CORRUPTED OBJECTS; **EER** REPRESENTS PERFORMANCE ON EVALUATION DATA, WHILE **HTER** REPRESENTS PERFORMANCE ON TEST DATA.

feature type / classifier	Approx. EER	HTER
moment inv. / gaus.	22.14	18.95
PCA / gaus.	0.00	3.19
HT / GMM	0.13	1.37

Occlusion experiments show that moment invariants are easily affected; this is due to the number of white pixels increasing (due to the presence of another object), and the center of mass now being in a different location. Both PCA and HT based approaches are relatively robust, though their performance still deteriorates as the overlap increases. The relative robustness of the PCA approach can be partly attributed to utilizing little or no information from the sides of the image, where the overlap is most pronounced. More formally, the elements, representing the sides of the image, in the eigenvectors of $\mathbf{U}$ in Eqn. (4), are relatively small compared to elements describing, say, the middle of the image; this is due to most training objects not extending all the way to the edges of the image (see Fig. 1 for an example). The relative robustness of the HT/GMM approach again stems from the parts based representation; the amount of corrupted blocks is dependent on the degree of overlap between the two objects; when the overlap is small, only a small subset of the feature vectors is affected.

5. CONCLUSIONS AND FUTURE WORK

We evaluated the performance of several combinations of feature extraction techniques and classification approaches for the purposes of classifying silhouettes in adverse conditions. The feature extraction techniques were based on Hu's moment invariants, Principal Component Analysis (PCA) and the 2D Hadamard Transform (HT); the first two methods are holistic in nature, while the last method is local in nature. The classifiers were based on gaussian densities and Gaussian Mixture Models (GMMs).

Out of the evaluated approaches, the HT/GMM approach seems the most promising; it is relatively robust to scale changes, clipping and occlusions. However, in its current form it is highly sensitive to noise or imaging artefacts, suggesting further research is required to address this limitation. For example, the post-processing heuristic could be improved by using a block occupation measure; only vectors, from blocks where the number of white pixels falls within upper and lower limits, could be used. Apart from ignoring blank or completely solid blocks, this would reduce training with noise-like patterns and passing many noisy patterns to the classifier.

The approach based on moment invariants and a gaussian classifier was shown to achieve relatively poor performance in advantageous conditions, and was easily affected by clipping and occlusions. The approach based on PCA features and a gaussian classifier was shown to be highly affected by scale changes and clipping, while being relatively robust to occlusions.

In terms of an operational system for finding a specific object based on its silhouette, there are several implications from the above observations. The performance of the system can be *highly dependent* on the performance of the Region of Interest (ROI) locator algorithm (i.e. the algorithm's ability to accurately locate an object, with no clipping or scale problems). As such, the pattern classification stage following

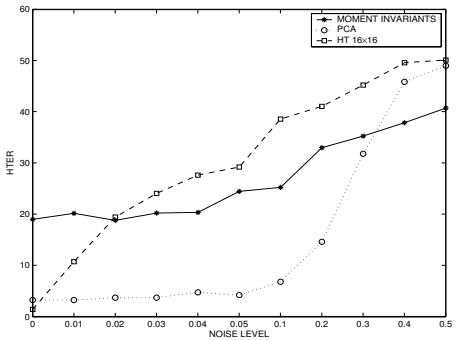

Fig. 4: HTER for noise corrupted objects.

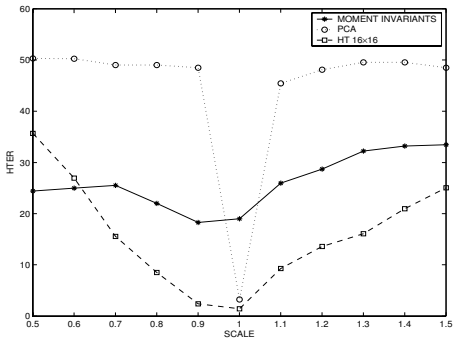

Fig. 5: HTER for scaled objects.

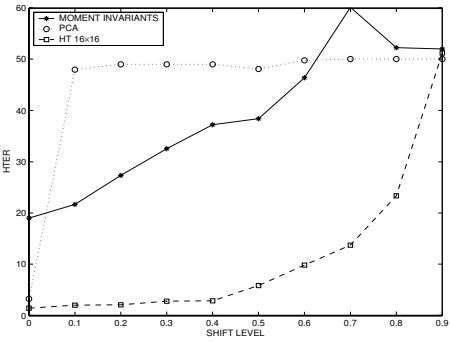

Fig. 6: HTER for clipped objects.

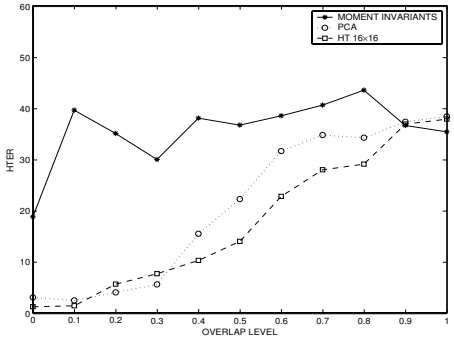

Fig. 7: HTER for occluded objects.

the ROI localization stage *must be able to handle imperfectly located ROI*.

We note that in this paper we have covered only a small subset of possible combinations of feature extraction techniques and classification approaches. Other notably popular approaches to classification include Artificial Neural Networks and Support Vector Machines [7], [30]; they were not considered here partly due to the small amount of object specific training data available in the present database. We also note that it is possible to create a GMM based system which utilizes holistic features; here, each training feature vector would be assumed to be a mean of a gaussian. This approach could model the distribution of training vectors more accurately than the gaussian based classifier. Other feature extraction techniques, such as the Fourier-Mellin Transform [11], [31] and Zernike moments [14], [31], could also be evaluated in future work.

ACKNOWLEDGEMENTS

This work was performed under the IRATD research agreement between CSSIP and DSTO. The authors thank M. Driscoll, M. Podlesak and T. Sills (DSTO, Edinburgh) for providing the object database. The authors also thank S. Searle (CSSIP) for useful suggestions.

REFERENCES

[1] S. Bengio et al., "Evaluation of Biometric Technology on XM2VTS", IDIAP Research Report 01-21, Martigny, Switzerland, 2001.

[2] F. Cardinaux, C. Sanderson, S. Bengio, "Face Verification Using Adapted Generative Models", In: *Proc. 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Seoul, 2004, pp. 825-830.

[3] A. Chan, S. Der, N. Nasrabadi, "Dualband FLIR fusion for automatic target recognition", *Information Fusion*, Vol. 4, No. 1, 2003, pp. 35-45.

[4] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Soc., Ser. B*, Vol. 39, No. 1, 1977, pp. 1-38.

[5] C.J.S. deSilva, G. Lee, R. Johnson, "All-aspect Ship Recognition in Infrared Images", In: *Proc. Electronic Technology Directions to the Year 2000*, Adelaide, 1995, pp. 194-198.

[6] G.R. Doddington et al., "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective", *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 225-254.

[7] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification*. 2nd Ed., John Wiley & Sons, USA, 2001.

[8] S. Dudani, K. Breeding, R. McGhee, "Aircraft Identification by Moment Invariants", *IEEE Trans. Computers*, Vol. 26, No. 1, 1977, pp. 39-45.

[9] J.-L. Gauvain, C.-H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. Speech and Audio Processing* Vol. 2, No. 2, 1994, pp. 291-298.

[10] R.C. Gonzales, R.E. Woods. *Digital Image Processing*, Addison-Wesley, 1992.

[11] A. Jin, D. Ling, O. Song, "An efficient fingerprint verification system using integrated wavelet and Fourier-Mellin invariant transform", *Image and Vision Computing*, Vol. 22, No. 6, 2004, pp. 503-513.

[12] B. Kamgar-Parsi et al., "Aircraft Detection: A Case Study in Using Human Similarity Measure", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 12, 2001, pp. 1404-1414.

[13] D.N. Kato et al., "Ship classification and aimpoint maintenance", In: *Infrared Systems and Components II*, Proc. SPIE, Vol. 890, 1988.

[14] A. Khotanzad, J.-H. Lu, "Classification of Invariant Image Representations Using a Neural Network", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 38, No. 6, 1990, pp. 1028-1038.

[15] A. Khotanzad, Y. Hong, "Invariant Image Recognition by Zernike Moments", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 5, 1990, pp. 489-497.

[16] B. Li et al., "Experimental Evaluation of FLIR ATR Approaches - A Comparative Study", *Computer Vision and Image Understanding*, Vol. 84, No. 1, 2001, pp. 5-24.

[17] S. Loncaric, "A survey of shape analysis techniques", *Pattern Recognition*, Vol. 31, No. 8, 1998, pp. 983-1001.

[18] S. Lucey, T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation", In: *Proc. IEEE Conf. Computer Vision and Pattern Recog.*, Vol. 2, 2004, pp. 855-861.

[19] D. Nair, J.K. Aggarwal, "Bayesian recognition of targets by parts in second generation forward looking infrared images", *Image and Vision Computing*, Vol. 18, No. 10, 2000, pp. 849-864.

[20] A. Pentland, B. Moghaddam, T. Starner, "View-Based and Modular Eigenspaces for Face Recognition", In: *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Seattle, 1994, pp. 84-91.

[21] J. Ratches et al., "Aided and Automatic Target Recognition Based Upon Sensory Inputs From Image Forming Systems", *IEEE Trans. Pattern Analysis and Machine Intell.*, Vol. 19, No. 9, 1997, pp. 1004-1019.

[22] D. Reynolds et al., "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Proc.*, Vol. 10, No. 1-3, 2000, pp. 19-41.

[23] S. Rizvi, N. Nasrabadi, "A modular clutter rejection technique for FLIR imagery using region-based principal component analysis", *Pattern Recognition*, Vol. 35, No. 12, 2002, pp. 2895-2904.

[24] S. Rizvi, N. Nasrabadi, "Fusion of FLIR automatic target recognition algorithms", *Information Fusion*, Vol. 4, No. 4, 2003, pp. 247-258.

[25] C. Sanderson, K.K. Paliwal, "Fast features for face authentication under illumination direction changes", *Pattern Recognition Letters*, Vol. 24, No. 14, 2003, pp. 2409-2419.

[26] S-G. Sun and H.W. Park, "Automatic target recognition using boundary partitioning and invariant features in forward–looking infrared images", *Optical Engineering*, Vol. 42, No. 2, 2003, pp. 524-533.

[27] M. Turk, A. Pentland, "Eigenfaces for Recognition", *J. Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86.

[28] G. K. Wallace, "The JPEG still picture compression standard", *IEEE Trans. Consumer Electronics*, Vol. 38, No. 1, 1992, pp. xviii-xxxiv.

[29] A. Waxman et al., "Neural Processing of Targets in Visible, Multispectral IR and SAR Imagery", *Pattern Recog.*, Vol. 8, 1995, pp. 1029-1051.

[30] A. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, UK, 2002.

[31] J. Wood, "Invariant pattern recognition: a review", *Pattern Recognition*, Vol. 29, No. 1, 1996, pp. 1-17.

[32] D. Zhang, G. Lu, "Review of shape representation and description techniques", *Pattern Recognition*, Vol. 37, No. 1, 2004, pp. 1-19.