ROC Method for the Evaluation of Multi-class Segmentation/Classification Algorithms with Infrared Imagery

G. S. Rees, W. A. Wright and P. Greenway
Advanced Information Processing Dept
ATC-Sowerby, BAE-SYSTEMS, PO-Box 5, Filton
Bristol, BS34 7QW, UK
gareth-s.rees@baesystems.com

Abstract

The classification of image regions of interest in an image is an important area of research. Generally most investigations concentrate on the optimisation of the constituent parts of the system without regard to the overall performance. This work takes a system centred approach. Using a novel multi-class receiver operating characteristic, which also allows for the inherent uncertainty present, it is shown that the influence of different region based segmentation algorithms on the performance of classification algorithms can be determined. The results generated, using this approach, for an airborne infrared application highlight the non-linear relationship between the constituent algorithms and show quantitatively that the system performance can be strongly class and segmenter/classifier dependent.

1 Introduction

The use of a segmentation, feature extraction, classification paradigm for labelling image regions is well established. This type of approach is well known for various applications including medical, visual, infrared, satellite and imaging radar, [6]. However, while there has been much work that considers the development and optimisation of the individual component parts there has been comparatively little, quantitative, work that considers the interaction of the different processing elements and their integration within a system. This is particularly important for subsystems that are embedded in a part of a larger system. Here it is essential to understand the relationship between the different components in the system. An important aspect of any systems analysis is to “balance” the performance of the individual components such that the system as a whole attains a desired performance within constraints of computational cost, robustness, etc.

A crucial requirement, necessary for a systems approach to succeed, is the easy characterisation of the performance of the individual components as a function of the overall system. However, performance characterisation is a difficult non-linear problem that can not be achieved by measuring the performance of individual components alone (see [5] Chap 11). It is necessary to allow for the effects of the processing elements that feed into any component. That is, better performance may be achieved through the use of a

BMVC 2002 doi:10.5244/C.16.52

matched segmentation/classification process rather than a one constructed from individual components that are optimal in their own right. To understand this issue it is necessary to obtain a measure of the performance of the overall system and compare this with the performance of the individual components. In the past such an analysis has been limited by the lack of available image data with which to conduct the study and the availability of useful measures of performance both for the segmentation and classification processes. Recent work, see the review in [11], has begun to consider the difficult problem of measuring segmentation performance in a well principled way and several measures currently exist. Castano et al [1] have also reported results of a study that tries to address these issues for classification problems. However, the results of this work were difficult to interpret. Performance is characterised using a raw confusion matrix, C_{ij} where rows (j) corresponding to "true" class labels and columns (i) corresponding to the labels estimated by the classifier, and the single summary statistic

$$P = \frac{Tr(C)}{\sum_{ij} C_{ij}}.$$

Although with this approach it is relatively easy to see the number of true positives and false negatives it is difficult to determine other performance measures, which may be more meaningful, such as the false positive. Furthermore, the summary statistic averages out important class dependent detail. An alternative approach that endeavours to remedy these difficulties is to derive Receiver Operating Characteristics (ROCs) [2]. Here the true positive rate is plotted on a graph against the false positive rate to give a clear pictorial view of the classifier performance. However, this approach is limited to *two* class problems. Ideally, therefore, what is needed is an ROC measure capable of dealing with multiple classes. Another component that has been overlooked, but is essential if a systems approach is to be taken, is to allow for the uncertainty in the performance estimates that are derived. This uncertainty arises from both the segmentation and classification processes. For the purpose of this paper we ignore the uncertainty derived from the variability of the segmentation process [11], and concentrate on that derived from the classifier.

In this paper we presents an extension of the ROC performance measure that allows for multiple classes together with the inherent uncertainty in the process. It is shown that these enhancements allow the detailed performance of a region based classification to be understood as a function of its constituent parts. By way of demonstration, the extended method is applied to the controlled analysis of a region classification system for a set of airborne infra-red images. This uses a number of segmentation and classification methods to determine the dependence of system performance on segmentation method, choice of region feature, and classifier type. The results generated using this extended ROC highlight the complex, non-linear, relationship between these components and quantitatively show that the system performance is strongly class and segmenter/classifier dependent.

2 Extended ROC Method

Typically the performance of any classifier will be a function of the *operating point* of the system (i.e. the classification threshold). It is often inappropriate to treat the cost of producing false positives (i.e. false alarms) as the same as the cost of producing false negatives (i.e., failures to detect). The use of Receiver Operating Characteristic (ROC) is often adopted to try and gain some insight into this problem [5]. For a given classifier system, a characteristic curve of the probability of detection p_d may be plotted as a function of the the false alarm rate p_f, given some classification threshold, on a class by class basis. Ideally a classifier will produce results that have a *high* p_d and *low* p_f, i.e., they approach the top left corner of the ROC plot. Poor performance on the other hand would have point close to the line $p_d = p_f$. See Figure 3 for examples of this representation. The extension of this idea to multiple classes is not a straight forward process. Typically the two class ROC curve generalises to an $n-1$ surface for n classes. A simpler approach, which is adopted here, is to choose a *single* operating point based on a risk/cost matrix that relates the cost of mis-classification (p_d, p_f) between all pairs of classes and then plot this on a class by class basis. For the basis of the results presented here both forms of mis-classification are treated equally. Thus

$$p_d(i) = \frac{C_{ii}}{\sum_j C_{ij}} \tag{1}$$

and

$$p_f(i) = \frac{\sum_j (C_{ji} - C_{ii})}{\sum_k (\sum_j C_{kj}) - \sum_j C_{ij}}. \tag{2}$$

If it is assumed that the distribution in p_d, p_f is Gaussian then to gain an indication of the uncertainty in the estimation of these values it is sufficient to calculate the mean $(\bar{p_d}, \bar{p_f})$ and variance (e_d, e_f) and plot error ellipse around the ROC value. Hand [5] describes methodologies for deriving the uncertainty in the ROC. The “more statistically powerful” is to take a Bayesian approach and average over a classifier’s parameters obtained by training on the same data. Alternatively one can estimate the variability of a single classifier using a number of different data sets. In the analysis presented here it is possible to undertake the first of these estimates for two of the classifiers, a Gaussian Mixture Classifier (GMMC) and MLP neural network. Ideally in both cases it would be desirable to undertake a full Monte-Carlo estimate of the means and variances. For the purpose of this study this was approximated using an extension of MacKay’s evidence approach [8] for the MLP. The GMMC approach used the average performance obtained classifiers derived from 200 randomly configured start points.

For the kNN classifier, used in this study, it is more difficult to take this approach. This is because for limited data, with uneven class frequencies, large values of k the classification is based toward the most frequent class. Consequently, it is difficult to average over the parameter k in a meaningful way. Hence for this classifier a binomial sample variance is assumed. Here

$$e_d^2(i) = \frac{\bar{p_d}(i)(1 - \bar{p_d}(i))}{\sum_j C_{ij}} \tag{3}$$

and

$$e_f^2(i) = \frac{\bar{p_f}(i)(1 - \bar{p_f}(i))}{(\sum_k \sum_j C_{kj}) - \sum_j C_{ij}}. \tag{4}$$

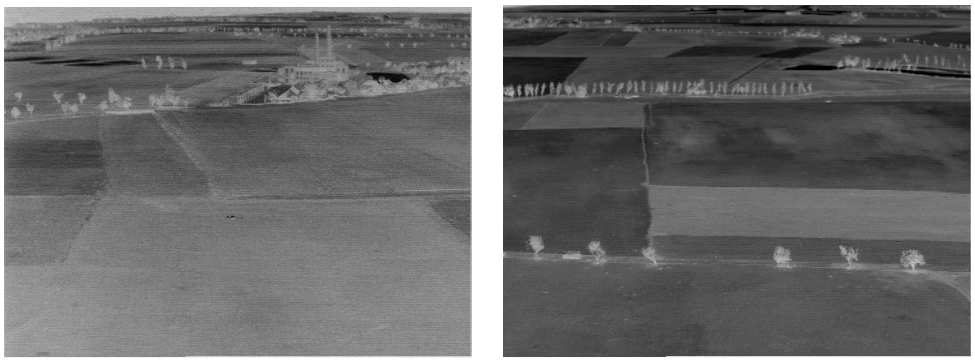


Figure 1: Typical Images

3 System Analysis for Infrared Region Classification

To demonstrate the utility of this multi-class ROC within a wider systems centred approach, it is used here to investigate the relative performance of a number of classifier/ segmentation combinations for an airborne infrared image classification problem. For this problem the classification system is broken down into a number of stages.

Image acquisition: A set of 150, 512×512, 3-5μ airborne infrared images was digitised (see typical images in Figure 1). Ground-truth, in terms of 9 classes, was then generated from these images using using a panel of human observers who hand labelled regions within the image. All results presented here were averaged across the entire image set.

Image segmentation: Two principle machine segmentation techniques were used for this study. The variational approach as developed by Mumford and Shah [9] and FORCE, a region agglomeration with an edge process as developed by Gay [4]. In addition to these, a control segmentation which was conceived to provide an indication of the upper performance bound on the segmentation process was employed. This "benchmark" segmentation was obtained directly from the classification ground-truth using label propagation. The selection operating points for the algorithms was undertaken using the M_4 performance measure [12]. This measure was applied to compare the output region maps and the benchmark segmentations as the segmentation parameters were varied. The algorithm parameters were then configured to yield maximum performance in M_4.

Feature generation and selection A variety of features were used for this work (see Table 1). The down selection and ranking of these features was achieved using a sequential backward search (SBS) [7]. This process was undertaken prior to training the classifiers. However, the performance criteria used to rank the features was based on the summary statistic P derived from a $(k = 8)$ kNN classifier [1] Again, the operating point for the

[1] It is recognised that the ideal is to use the same technique for feature selection as for actual classification. However, this was not considered necessary for the purpose of this study.

Feature	**Benchmark**	**FORCE**	**Mumford-Shah**
1 Size			
2 Log size	3		5
3 Brightness	8	4	*3*
4 Brightness Sd	6		13
5 HE brightness	*1*	*3*	
6 HE brightness Sd			
7 Vertical position	4	*2*	*2*
8 Horizontal position		8	7
9 Cert. pos. Sd			
10 Horiz. pos. Sd			15
11 Elongation	10	9	
12 $\mathbf{L}_0$ texture	12	13	
13 $\mathbf{L}_1$ texture			6
14 $\mathbf{L}_2$ texture		7	11
15 $\mathbf{L}_3$ texture			
16 $\mathbf{L}_4$ texture	*2*	12	14
17 $\mathbf{L}_5$ texture	9	5	10
18 $\mathbf{L}_6$ texture			9
19 $\mathbf{L}_7$ texture		11	12
20 $\mathbf{L}_8$ texture	7	6	4
21 2nd order invariant ψ_1	5		
22 2nd order invariant ψ_2	11	10	8
23 3rd order invariant ψ_3			
24 3rd order invariant ψ_4			
25 3rd order invariant ψ_5			
26 3rd order invariant ψ_6			
27 3rd order invariant ψ_7		*1*	*1*
28 Orientation			

Table 1: Features and rankings from down-selection. The numbering indicates the feature ranking for each segmentation method with 1 being the most important. The top three features for each segmentation method and, hence, salient features are shown in *italic*. These features include Histogram Equalised Features (HE), Laws textures (L) and image invariants (ψ).

subsequent stages was obtained by removing uninformative features while system performance increased to a maximum. Figure 2 shows the variation in performance during the SBS process while Table 1 summarises the rankings for those features that were retained in the reduced feature sets.

Classifiers: Three typical forms of classifiers were used for this study.

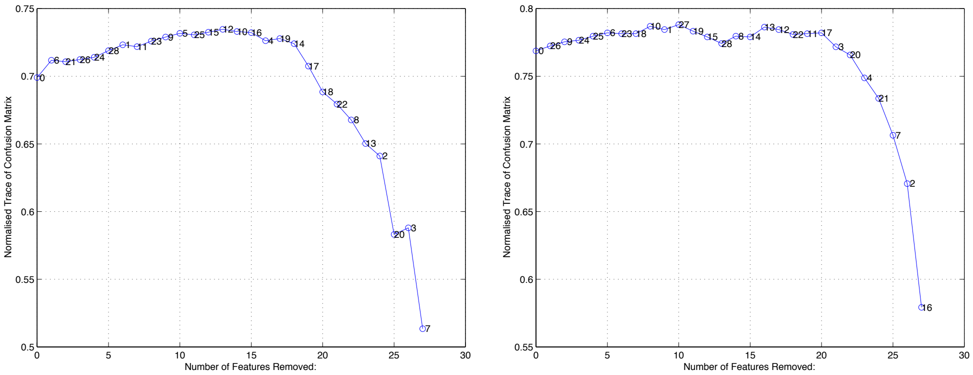
A *voting* kNN algorithm was used for the non-parametric classifier. This used a Euclidean distance measure to rank the k nearest members of the training set. The value of k was obtained using a leave one out cross-validation strategy for $k = 3, 5, \ldots, 9$ for the reduced feature set. The optimal value of k was found to be $k = 8$.

The GMMC explicitly modelled the class posterior probability for the feature vector (or data) $\boldsymbol{x}$ given some training data D as

$$p(\mathcal{C}_i|\boldsymbol{x}, D) = \frac{p(\mathcal{C}_i)\, p(\boldsymbol{x}|\mathcal{C}_i, D)}{p(\boldsymbol{x})}, \tag{5}$$

where $p(\mathcal{C}_i)$ is the prior over class i and $p(\boldsymbol{x}|\mathcal{C}_i, D)$ is termed the class conditional likelihood. This likelihood is modelled mixture of kernels

$$P(\boldsymbol{x}|\mathcal{C}_i) = \sum_j K((\boldsymbol{x} - \boldsymbol{x}_{ij}); \sigma_{ij}^2)$$



(a) Mumford-Shah Segmentations

(b) Benchmark Segmentation

Figure 2: Feature Saliency from SBS for features derived from Mumford-Shah and Benchmark Segmentations

where

$$K((\boldsymbol{x}-\boldsymbol{x}_{ij});\sigma_{ij}^2) = p(\boldsymbol{x}_{ij}|D)\,\mathcal{N}(\boldsymbol{x}_{ij},\sigma_{ij}^2),$$

$\mathcal{N}(\boldsymbol{x}_{ij},\sigma_{ij}^2)$ is a Gaussian centred about some point $\boldsymbol{x}_{ij} \in \mathcal{X}_i$ with variance σ_{ij}^2, $\mathcal{X}_i$ is the set of all kernel centres belonging to class i and $p(\boldsymbol{x}_{ij}|D)$ weights the relative importance of each centre j for class i [13]. This allows Equation 5 to be written as:

$$P(\mathcal{C}_i|\boldsymbol{x},D) = \frac{p(\mathcal{C}_i)\sum_j K((\boldsymbol{x}-\boldsymbol{x}_{ij});\sigma_{ij}^2)}{p(\boldsymbol{x})}, \tag{6}$$

where denominator

$$p(\boldsymbol{x}) = \sum_i p(\mathcal{C}_i)\sum_j K((\boldsymbol{x}-\boldsymbol{x}_{ij});\sigma_{ij}^2),$$

provides the appropriate normalisation.

The classification of the feature vector $\boldsymbol{x}$ is then obtained by thresholding equation 6 using a “winner takes all” strategy which treats all mis-classifications equally. The centres $\boldsymbol{x}_{ij}$ for the kernels was initiated using a k-means algorithm. Fitting of the kernels was undertaken by applying the EM algorithm [3], this re-estimates the mean and variance of the Gaussian kernels, together with the weighting probability $p(\boldsymbol{x}_{ij}|D)$. The improper prior

$$p(\mathcal{C}_i) = 1/f_i$$

was used for the class prior where f_i is the frequency of the i^{th} class in the training data.

The advantage of this method is that it uses a separate set of kernels for each class. These are tuned using data only from that class. The method, therefore, allows for different frequencies of for each class in the training data and so is ideally suited to classification problems where the number of training examples varies widely between classes as in the case explored in this paper.

An MLP classification neural network was used to provide a hyper-parametric classifier. For this study a three-layer network architecture was adopted. This had a single hidden layer and an n node output layer where each node corresponded to a particular class $\mathcal{C}_i$. The weights of the network were estimated using an extension of, Bayesian, evidence approach [8] as described for multiple classes by Nabney [10]. Here a quadratic weight regularisation term was introduced during training and Softmax activations functions used in the output layer. This ensures that the i^{th} output of the network gives the class conditional probability $p(\mathcal{C}_i|\boldsymbol{x}, D)$. As with the GMM classifier the classification was then obtained by "winner takes all" thresholding of the posterior. Importantly, this Bayesian approach also provides a Gaussian based estimate the uncertainty in the class conditional probability, which in turn was used to estimate the uncertainty in the ROC plot.

To determine the number of hidden units the data was split ($75\%/25\%$) into training and validation data. For each segmentation method the performance on the validation set was then monitored for differing number of hidden units. The network that gave the best performance was then selected and final classification results generated.

Segmentation Method:	**Benchmark**	**FORCE**	**Mumford-Shah**
kNN	76.1 ± 0.4	73.6 ± 0.2	73.3 ± 0.3
GMM	81.5 ± 0.8	71.2 ± 0.4	71.8 ± 0.6
MLP	76.5 ± 0.8	73.8 ± 0.4	69.6 ± 0.6

Table 2: The effect of the choice of segmentation/classification method on system performance. The table shows the overall performance results obtained from the three different classifiers for the reduced feature set using three different segmentation methods. Each entry is the estimated overall mean percentage of regions correctly classified in each experiment. Errors are at the 1-sigma level and estimated as described in the text.

4 Results

Figure 3 shows the new multi-class ROC plots for six out of the nine possible combinations described in Section 3. Each pane within the figure shows a plot of results from an experiment with one segmenter classifier combination.

In each plot the y-axis is the true-positive probability, p_d, while the x-axis is the false-positive probability, p_f. The mean performance for each class are plotted as solid points using a unique colour as shown in the legends. The classes considered were: field type 1 (bare), field type 2 (with crops), copse/wood/forest, hedge-row, road, town or built-up areas, individual building, sky and water. The ellipses about each plotted point indicate the uncertainty estimated for that class using the results of the experiment from 150 test images. Notice that these uncertainty ellipses are generally elongated along the y-direction indicating that the p_d was found to be more uncertain than the p_f.

Clearly, a good system would have results for all classes that cluster in the top-left of each plot, signifying high p_d and low p_f. In contrast, a system with points close, or whose error ellipses intersect, the diagonal line $p_d = p_f$ would be "guessing" and have little utility. This diagonal lower limit on performance is shown as the hollow points, coloured appropriately for each class. These indicate the classification performance that would be obtained by assigning class using only the measured frequency for image regions belonging to that class.

It is apparent from these results that there is significant variation in performance both between individual classes and between combinations of segmenter and classifier. It can be seen that almost all of the results are significantly better that that would be obtained by assigning classes on the basis of measured frequency. However, in general, exceptions were obtained from combinations involving the kNN and MLP classifiers (see typical examples are shown in panes d, e and f). Here classes field type 1 and copse/wood are close to the diagonal. Also, was also noticed that most results, regardless of segmenter/classifier choice, are clustered close to the y-axis indicating low false positive rates. Exceptions were found for combinations involving the FORCE segmenter (a typical example is shown in pane b), where field type 2 exhibits excessive false positives. Preliminary investigations suggest that is is due to the FORCE segmentation merging fragments of field into neighbouring regions.

Comparisons based on the choice of classifier type suggests that the GMMC generally performs both better and more consistently than the other classifiers. Class by class results can be seen to cluster at higher values of p_d and low values p_f across the range of segmenters. In contrast both the kNN and MLP results were found to be generally lower in p_d, were more spread out and exhibited a far greater sensitivity to the choice of segmenter. Comparison of the results from the kNN classifier combined with the Benchmark and Mumford-Shah segmenters show that better performance was obtained from the machine segmentation rather than the Benchmark. It is speculated that this is due to the relatively larger numbers of regions obtained from the Mumford-Shah algorithm. This result highlights the complex interaction between segmentation and classification components.

Table 2 summarises the results measured using the summary statistic P. The contrast between this *averaged* performance measure and the extended Multi-Class ROC is clear; the class averaged summary statistic results show little variation across the nine system configurations and so is not as informative. From these observations it is clear that this extended multi-class ROC provides insight into the relative effects of the classifier and segmenter subsystems.

5 Conclusions

In this paper multi-class extension to the ROC which also allows for uncertainty has been demonstrated on an infrared-image region-classification application. It has been shown that this ROC method allows insight into the complex relationship between between segmentation and classifier components of a region classification system. The results also show, at least for the implementations of the three classifiers and two segmenters considered, that performance strongly depends on classifier/segmenter pairing; a difference that is not apparent from a simple summary statistic. Furthermore, a very significant class by class variation is also apparent. Lastly, although the function of this study is not to compare segmentation and classification methods, they suggest that for the imagery considered a GMMC offers the both the best and most consistent performance and not surprisingly that improved segmentation may yield significant improvements to system performance.

References

[1] R. et Al Castano. Classification experiments on real-world texture. *CVPR-2001: Proceedings of Worksop on Empirical Methods in Computer Vision*, pages 1–22, 2001.

[2] P. Courtney and N.A. Thacker. Performance characterisation in computer vision: The role of statistics in testing and design. In Blanc-Talon and Popescu, editors, *Imaging and Vision Systems: Theory, Assessment and Applications*. NOVA Science Books, 2001.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

[4] M. Gay. Segmentation using region merging with edges. In *Proceedings of the Third Alvey Vision Conference*, pages 115–119. 1989.

[5] D.J. Hand. *Construction and Assessment of Classification Rules*. John Wiley, 1997.

[6] IEEE. *29th Applied Imagery Pattern Recognition Workshop (AIPR'00)*. IEEE Press, 2000.

[7] J. Kittler. Feature selection and extraction. In T. Young and K Fu, editors, *Handbook of Pattern Recognition and Image Processing*, pages 59–83. NY Academic Press, 1986.

[8] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

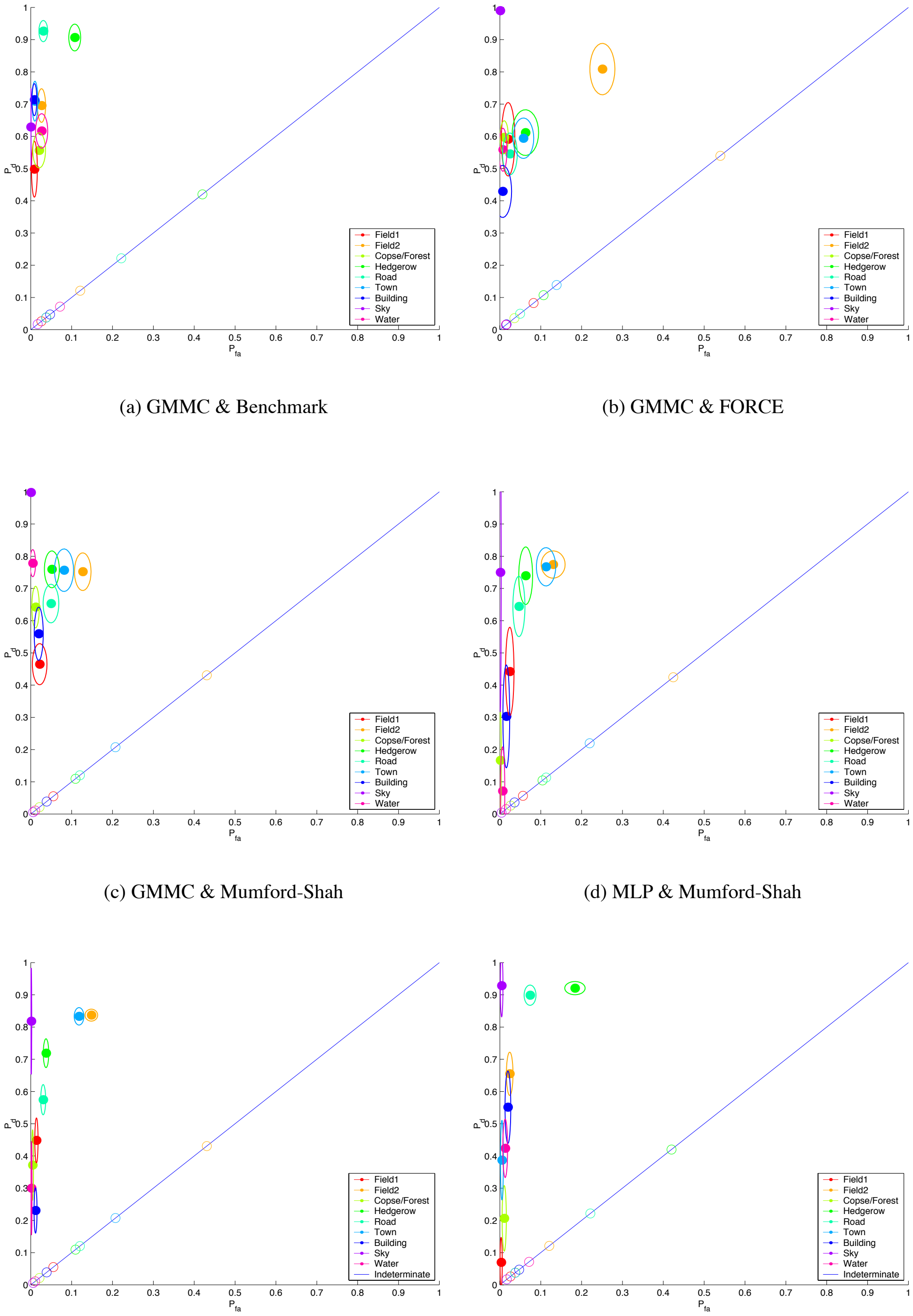
[9] D. Mumford and J. Shah. Boundary detection by minimising functionals. *Image Understanding*, 1988.

[10] I.T. Nabney. *NETLAB: Algorithms for Pattern Recognition*. Springer, 2001.

[11] G. Rees. *A framework for the development of applications involving image segmentation*. PhD thesis, Oxford Brookes University, 1987.

[12] G. Rees and P. Greenway. Metrics for image segmentaion. In *Proceedings of the workshop on performance evaluation, International Conference Vision Systems (ICVS 99)*. 1999.

[13] B.W. Silverman. *Density Estimation*. NY Chapman and Hall, 1986.



(a) GMMC & Benchmark

(b) GMMC & FORCE

(c) GMMC & Mumford-Shah

(d) MLP & Mumford-Shah

(e) kNN & Mumford-Shah

(f) kNN & Benchmark

Figure 3: Muli-Class ROC plots as a function of segmenter and classifier for the reduced feature set.