On the Performance Characterisation of Image Segmentation Algorithms: A Case Study

B. Southall[1,2,3], B.F. Buxton[2], J.A. Marchant[3], and T. Hague[3]

[1] GRASP Laboratory, University of Pennsylvania,
3401 Walnut Street,
Philadelphia, PA 19104, USA
southall@grip.cis.upenn.edu
Tel: +1 215 898 0352 Fax: +1 215 573 2048

[2] Department of Computer Science, University College London, Gower Street,
London, WC1E 6BT, UK
b.buxton@cs.ucl.ac.uk
Tel: +44 20 7679 7294 Fax: +44 20 7387 1397

[3] Silsoe Research Institute, Wrest Park, Silsoe,
Bedfordshire, MK45 4HS, UK
{john.marchant,tony.hague}@bbsrc.ac.uk
Tel: +44 1525 860000 Fax: +44 1525 860156

Abstract. An experimental vehicle is being developed for the purposes of precise crop treatment, with the aim of reducing chemical use and thereby improving quality and reducing both costs and environmental contamination. For differential treatment of crop and weed, the vehicle must discriminate between crop, weed and soil. We present a two stage algorithm designed for this purpose, and use this algorithm to illustrate how empirical discrepancy methods, notably the analysis of type I and type II statistical errors and receiver operating characteristic curves, may be used to compare algorithm performance over a set of test images which represent typical working conditions for the vehicle. Analysis of performance is presented for the two stages of the algorithm separately, and also for the combined algorithm. This analysis allows us to understand the effects of various types of misclassification error on the overall algorithm performance, and as such is a valuable methodology for computer vision engineers.

1 Introduction

Economic and ecological pressures have led to a demand for reduced use of chemical applicants in agricultural operations such as crop and weed treatment. The discipline of precision agriculture strives to reduce the use of agro-chemicals by directing them more accurately and appropriately. The extreme interpretation of this approach is *plant scale husbandry*, where the aim is to treat individual plants according to their particular needs. An experimental horticultural vehicle has been developed to investigate the viability of plant scale husbandry, and

D. Vernon (Ed.): ECCV 2000, LNCS 1843, pp. 351–365, 2000.

previous work [15,17,16] has described a tracking algorithm, centred upon an extended Kalman filter, that allows navigation of the vehicle along the rows of crop in the field. This paper presents a simple algorithm for frame-rate segmentation of images for the task of differential plant treatment, together with a thorough evaluation of algorithm performance on data captured from the vehicle. The algorithm comprises two stages. Stage I aims to extract image features which represent plant matter from the soil background, and stage II divides these features into crop and weed classes for treatment scheduling.

The practical application of the algorithm requires that we understand how its performance varies in different operating conditions; Haralick [10] underlines the necessity of the evaluation of computer vision algorithms if the field is to produce methods of practical use to engineers. In this paper, we evaluate the two stages of the algorithm separately and as a result, we are able to gain deeper insight into the performance of the algorithm as a whole. A review of techniques for image segmentation evaluation is presented by Zhang [22], who partitions the methods into three categories; analytical, where performance is judged on the basis of its principles, complexity, requirements and so forth; empirical goodness methods, which compute some manner of "goodness" function such as uniformity within regions, contrast between regions, shape of segmented regions; and finally, empirical discrepancy methods, which compare properties of the segmented image with some ground truth segmentation and computes error measures. Analytic methods may only be useful for simple algorithms or straightforward segmentation problems, and the researcher needs to be confident of the models on which these processes are based if they are to trust the analysis. Empirical goodness methods have the advantage that they do not force the researcher to perform the onerous task of producing ground truth data for comparison with the segmentation, for meaningful results, an appropriate model of "goodness" is required, and in most practical problems if such a model were available, it should be used as part of the algorithm itself. This leaves empirical discrepancy methods, which compare algorithmic output with ground truth segmentation of the test data and quantify the levels of agreement and/or disagreement.

A discrepancy method which is suitable for two-class segmentation problems is receiver operating characteristic (ROC) curve analysis. Rooted in psychophysics and signal detection theory, ROC analysis [8,21] has proved popular for the comparison of diagnostic techniques in medicine [11,9], and is gradually gaining currency within the computer vision and image analysis community for the comparative evaluation of algorithms such as colour models [2], edge detectors [1,5] and appearance identification [6]. Receiver operating characteristic curves typically plot true positive rates against false positive rates as a decision parameter is varied and provide a means of algorithm comparison and evaluation. ROC analysis also allows selection of an operating point which yields the minimum possible Bayes risk [8]. ROC curves will be discussed further below, together with the related maximum realisable ROC (MRROC) curve [14]. Although our algorithm produces a three-way final classification (crop, weed and soil), stages

I and II are both binary classifiers, so we can analyse their performance using ROC methods.

We will first outline the segmentation algorithm prior to discussing evaluation of the performance of stages I and II. Final results for the complete algorithm are then presented and discussed in the light of our knowledge of its constituent parts.

2 The Segmentation Algorithm

The two stage segmentation algorithm is sketched in the following sections; for the sake of brevity, details of the algorithms are not given (these may be found elsewhere [15,18]), but sufficient information is provided to allow the performance evaluation sections to be understood.

2.1 Stage I: Plant Matter Extraction

The experimental vehicle is equipped with a monochrome camera that is fitted with a filter which blocks visible light, but allows near infra-red wavelengths to pass. Many researchers, including for example Biller [4], have noted that the contrast between soil and plant matter is greater in the near infra-red wavelengths than the visible, and this allows us to use a grey level threshold to extract pixels which represent plant matter from the images captured by the vehicle as it traverses the field. We use an adaptive interpolating threshold algorithm, to allow for the fact that there is often a brightness gradient across many of the images captured by the vehicle. The cause of such a gradient is most likely the position of the Sun relative to the ground plane and the vehicle's camera, and the interaction of the illuminant with the rough surface of the soil. A simple linear variation in intensity between the upper and lower parts of the image is used to allow for such effects. Accurate modelling of illumination and reflectance effects is a complex issue and not of direct concern to this work. More principled models are known for surface reflectance, such as those due to van Branniken *et al* [20] or Oren and Nayar [12].

The algorithm is also adaptive to the average brightness of the image, which offers some robustness to changes in illumination as, for example, when the Sun is temporarily masked by a cloud. A mean grey-level is computed for both the top (μ_1) and bottom (μ_2) halves of the image and these two means are used as fixed points to linearly interpolate a mean $\mu(y_f)$ across the vertical pixel coordinates of the image. The classification of output pixels $O(x_f, y_f)$ is then given by the adaptive interpolating thresholding algorithm:

$$O(x_f, y_f) = \begin{cases} \text{P if } I(x_f, y_f) \geq \alpha\mu(y_f) \\ \text{S if } I(x_f, y_f) < \alpha\mu(y_f) \end{cases}, \tag{1}$$

where P denotes plant matter (crop or weed) and S soil. The decision rule of equation 1 is used in a chain-code clustering algorithm [7] whereby groups of neighbouring above-threshold pixels are clustered into "blobs". Each blob is

described by the pixel co-ordinates of its centroid in the image, and its size in number of pixels. The process is illustrated in figure 1 which shows an image and the plant matter extracted from it automatically. It can be seen from the figure that some of the plants fracture into multiple blobs. This is largely caused by shadows falling between plant leaves which lead to areas of the plant in the image that lie below the chosen threshold. Another problem that sometimes occurs is that neighbouring plants sometimes become merged into a single feature. Whilst there is little that can be done about the latter problem, the feature clustering technique in stage II of the algorithm aims to address difficulties caused by plant features fracturing.

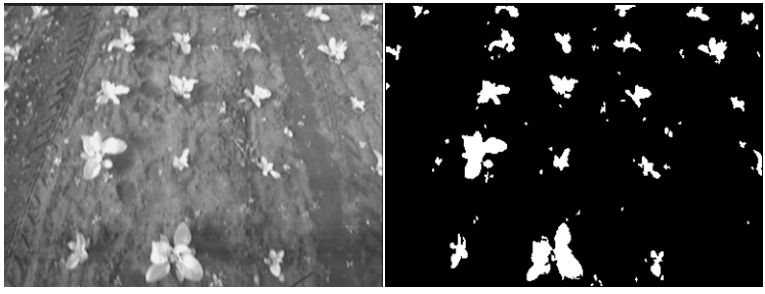

Fig. 1. An image and its automatically extracted plant matter.

2.2 Stage II: Crop/weed Discrimination

The image on the left of figure 1 shows that the crop plants grow in a fairly regular pattern in the field, and also that they are generally larger than the weeds. These are the two pieces of information that we exploit in the second stage of the segmentation algorithm, which aims to separate the set of plant matter features (denoted P) into subsets of crop (C) and weed (W). The first step of this stage is to filter the plant matter features on the basis of their size in the image. Justification for this decision is provided by figure 2, where histograms of the feature sizes (in pixels/feature) are plotted for both weed and crop. This data is derived from manually segmented images that we use as our ground truth data throughout this paper. More details of this data are given below.

It can be seen from the histograms that the vast majority (in fact 95%) of the weed blobs have a size of less than 50 pixels, whilst most (90%) of the crop blobs have a size of 50 or pixels or greater. This supports the claim that the weeds are typically smaller than the crop.

Thus, we have a straightforward algorithm that places a threshold on the size s of the image features. This may be expressed as follows:

$$\text{Class(feature)} = \begin{cases} W \text{ if } s(\text{feature}) < \varsigma \\ P \text{ if } s(\text{feature}) \geq \varsigma \end{cases}, \tag{2}$$

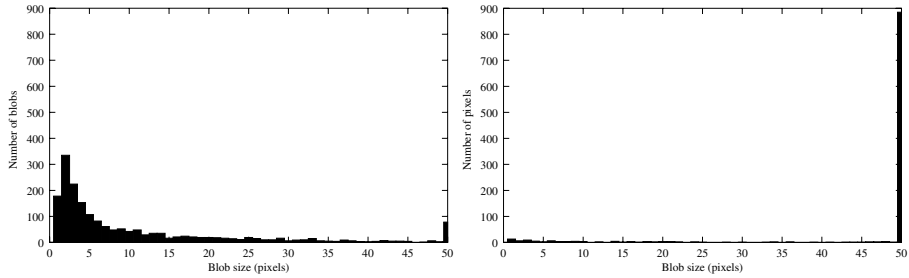

Fig. 2. Blob size histograms. Left: weed blobs. Right: crop blobs. In both histograms, the right-most bin (marked 50) counts all blobs of size ≥ 50. Note that the crop feature histogram, most of the bins are empty, except for the right-most.

where s(feature) is the size of an image feature in pixels, and ς is the size threshold.

The second step of stage II of the algorithm makes use of the regular grid pattern formed by the crop as they are planted in the field. The grid pattern is used as a cue for vehicle navigation [15], where the position of the vehicle relative to the crop grid and the dimensions of the grid are estimated by an extended Kalman filter (EKF) [3]. The EKF also produces a covariance matrix that describes the level of confidence in the current estimate. The state estimate is used to predict the position of each plant within the grid, and an algorithm akin to a validation gate [13] is used to cluster all plant matter features within a certain radius of the predicted crop plant position.

The validation gate has proved to be effective as an outlier rejection mechanism in practical Kalman filtering applications [13]. The algorithm combines the uncertainty on the predicted feature position and the uncertainty attached to the observed data to define a validation region outside of which candidate feature matches are rejected as being outliers. In our algorithm, we take the uncertainty of the estimated plant position and combine it with a user defined region which describes a radius on the ground plane about the plant centroid within which all of the crop plant should lie. This defines an association region in the image inside of which all plant matter features are labelled as crop (C), and outside of which the features are labelled as weed (W). The schematic diagram in figure 3 illustrates the components of the association region, Full details of the algorithm can be found elsewhere [18]. The size of the region which describes the user-defined plant radius is controlled by a single parameter r, the radius on the ground plane within which the crop plant matter should lie. This model implicitly assumes a distribution for the weed matter that gives lower probability of weed occurrence than plant occurrence within the radius r.

3 Evaluation Using ROC Curves

The receiver operating characteristic (ROC) curve [8] supports the analysis of binary classification algorithms whose performance is controlled by a single para-

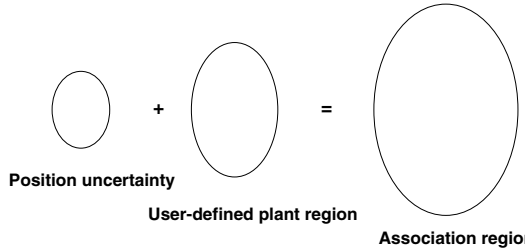

Fig. 3. The construction of the association region.

meter. For each parameter setting, algorithmic output is compared with ground truth data, and four numbers are calculated; TP, the number of "positive" cases correctly classified as positive; TN, the number of "negative" cases correctly classified as negative; FN, the number of positive cases incorrectly classified as negative; and FP, the number of negative cases incorrectly classified as positive. In the statistical literature, FN cases are type I errors, and FP cases type II errors [19]. From these four figures, two independent quantities are constructed, the true positive ratio, TPR = TP/(TP+FN), and the false positive ratio, FPR=FP/(FP+TN). To construct an ROC curve, a set of algorithm parameter values are chosen, and for each of these, the TPR and FPR values are calculated and plotted against each other. The set of TPR,FPR pairs form the ROC curve.

To characterise the performance of our algorithms, we shall use the area underneath the ROC curve. This metric has often been used to compare the performance of different algorithms across the same data sets [2,1,6], but we will use it to compare the performance of stage I of our algorithm across a number of test data sets which represent different stages of crop growth and weather conditions that the vehicle is likely to encounter. The performance of stage II across these data sets is assessed using the maximum realisable ROC (MRROC) curve. It is also possible to use the slope of the ROC curve to select a value for the algorithm's controlling parameter which minimises the Bayes risk associated with the decision being made. van Trees [21] provides full details.

3.1 The MRROC Curve

In the analysis described above, variation of a single decision parameter in a classification algorithm leads to the formation of the ROC curve. Each point on the curve characterises an instance of the classification algorithm that we call a *classifier*. If a single parameter is undergoing variation, then all of the classifiers lie along the ROC curve. This is the case within stage I of our algorithm, the adaptive interpolating threshold, which has gain parameter α, as defined in equation 1.

When an algorithm has more than one parameter, then it will generate a cloud of classifiers in the ROC space. The convex hull of this cloud is the MRROC curve [14], and the area underneath it reflects the best overall classification performance it is possible to obtain from this group of classifiers. We will use

the area under the MRROC curve to compare the operation of stage II of our algorithm, which has two parameters ς and r (the size threshold and clustering radius, respectively), on different data sets. The set of classifiers which provide the best performance is comprised of those that lie on the MRROC curve. Unlike the normal ROC curve which is a function of one decision parameter alone, it is not possible to set the algorithm operating point on the basis of the slope of an MRROC curve.

4 Characterisation of the Algorithm

We will now deal with algorithm characterisation, which is the evaluation of algorithmic performance over a range of different data sets. For the purposes of performance evaluation, we require image data sets which are representative of the application, and also a set of labelled images which represent the "true" segmentation of these scenes into the classes of interest to compare with the algorithmic output [22].

4.1 Ground Truth Image Data

Four sequences of images captured from the vehicle were used in off-line tests of the classification algorithm. An example image from each sequence is given in figure 4 (a)–(d). The sequences have been chosen to represent a range of typical crop growth stages and imaging conditions, although this range should by no means be considered exhaustive. The sequence properties are summarised in table 1. The deep shadows seen in figure 4 D are a result of bright sunlight.

Sequence	# images	Crop age	Weed density	Weather
A	960	8 weeks	low	cloudy
B	960	3 weeks	very low	overcast
C	1380	6 weeks	moderate	overcast
D	1280	3 weeks	very low	sunny

Table 1. Properties of the image sequences.

Haralick [10] asserts that performance characterisation requires a test set of statistically independent data. To this end, a subset of each image sequence was chosen such that no two images contain overlapping areas of the ground, which ensures that no two pixels in the test set represent the same patch of soil or plant. For each image in these test sets (a total of 66 images across the four sequences), a ground truth labelling was produced by hand segmenting the image pixels into four classes: crop, weed, soil and doubt. The ground truth images have been produced by hand using standard image editing software, and are subject to error, especially at border pixels where different image regions (crop, weed or soil) are adjacent. Some of these pixels will be incorrectly classified as their adjacent class, whilst some will be of genuinely mixed class. Alexander [2] noted such problems with border pixels and proposed that at the border between foreground (in our case plant matter) and background (soil), regions of doubt

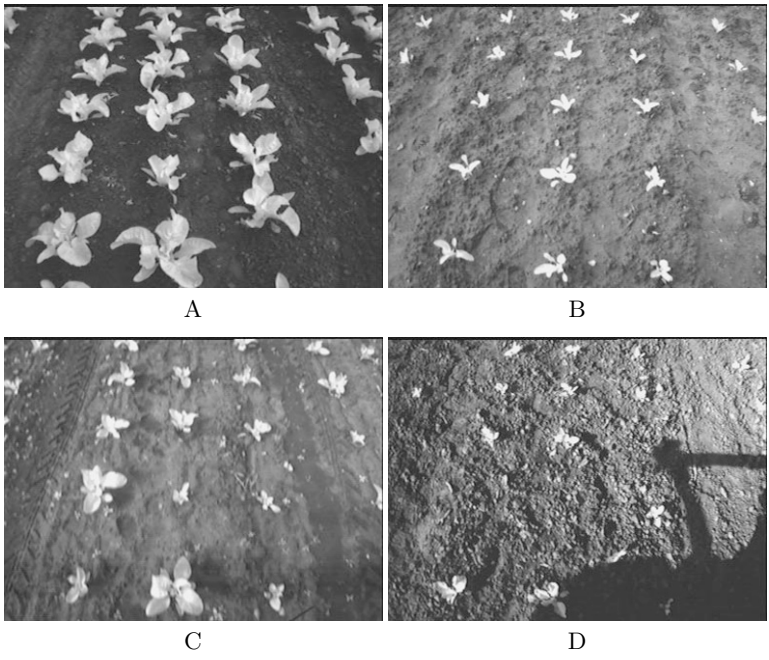

Fig. 4. Examples from the four image sequences A – D.

should be inserted, and the pixels within these doubt regions should be ignored for the purpose of assessing classifiers. All pixels that are on the border of plant matter and soil in the ground truth images are assigned to the doubt class and ignored in the classification assessments.

4.2 Stage I

A set of 27 threshold gain levels was chosen and the algorithm applied to the test images to generate the TPR,FPR pairs that constitute the ROC curve. The area under each of the curves plotted for sequences A – D is given in table 2.

Data Set	Area under ROCC	Area under MRROCC
A	0.9957	0.9974
B	0.9779	0.9997
C	0.9846	0.9996
D	0.9241	0.9993

Table 2. Area underneath ROC curves for algorithm stage I, sequences A–D (left) and for the MRROC curves for algorithm stage II (right).

The performance of stage I on each of the four data sets is reflected by the measures of area underneath the ROC curve shown in table 2. These show that the algorithm performs best on sequence A, where the plants are large and

there are few shadows, with sequences C and B following. The lowest overall performance is seen on sequence D, caused by the heavy shadows present (figure 4 D).

4.3 Stage II

As noted above, to compare the performance of stage II of the algorithm on our data sets, we use MRROC analysis. The two parameters ς and r, described in section 2.2, are varied systematically (27 samples of each parameter, yielding a total of 729 classifiers) to produce a cloud of TPR,FPR pairs in the ROC space. The convex hull of these points constitutes the MRROC curve [14], and the area underneath the curve is calculated for each data set and used as a metric for comparison, with better performance indicated as usual by an area closer to 1.

Recall that stage II of the algorithm comprises two steps, a size filtering step followed by feature clustering on the basis of proximity to the crop grid pattern. In the fully automatic algorithm, the input features are provided by stage I, and the grid position by the extended Kalman filter crop grid tracker [15]. In our first experiment, we removed the dependency on both of these algorithms by locating the crop grid by hand, and used the ground truth classified features as our input. In his outline of a performance characterisation methodology Haralick [10] states that testing algorithms on perfect input data is not worthwhile; if the algorithm's performance is less than perfect, then a new algorithm should be devised. In an ideal world, this would be the case, but our the crop/weed discrimination problem is difficult; capturing the large variations in size and shape of each sort of plant devising an algorithm to fit such models to image data will not be easy, so we currently have to settle for an imperfect algorithm that makes mistakes even on perfect data. In this case, testing on perfect input data tells us the best performance that the algorithm can be expected to deliver.

The areas underneath the MRROC curve for each sequence in this experiment are given in the right-hand columns of table 2, whilst a section of the MRROC curve, and the cloud of classifiers in the ROC space, is plotted in figure 5 (where we take crop pixels to be positives and weed pixels to be negatives). In table 2, the performance of the stage II algorithm is seen to be consistent over each sequence, and very close to the ideal of 1 in each case. As noted above, to generate the curve for each sequence, we ran 729 trials of the algorithm over each of the 4 sequences, a time-consuming task. To cut down on computational effort for the fully automatic algorithms, we selected a single ς, r pair for each sequence. The point selected was that closest to the ideal (0,1) point in ROC space. A more principled selection of operating parameters might be possible if the values and costs of correct and incorrect decisions were known. For example, if the farmer wishes to remove all weeds and is willing to risk some crop in this process, the value of true negatives (correctly classified weed) would be high, and the cost of a false positive (weed classified as crop) would be higher than the cost of a false negative (crop classified as weed). If crop fertilisation was a priority, a true positive (correctly identified crop) would be high, and the cost of a false negative would be higher than the cost of a false positive. The values of ς and r chosen,

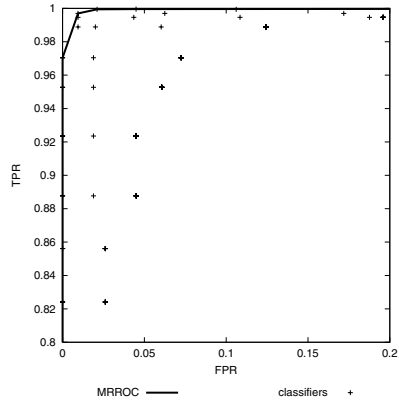

Fig. 5. The MRROC curve for ground truth plant matter segmentations of sequence C.

together with their corresponding TPR and FPR, are given for each sequence in table 3.

			Parameter Setting		Automatic Tracking	
Sequence	ς (pixels)	r (mm)	TPR	FPR	TPR	FPR
A	100	450	0.9950	0.0	0.9939	0.1564
B	30	100	0.9940	0.0	0.9982	0.0
C	80	100	0.9970	0.0094	0.9981	0.0307
D	30	100	0.9975	0.0638	0.9993	0.2017

Table 3. Operating points for the size filtering and clustering algorithms, and their corresponding TPR and FPR chosen in the parameter setting experiment (left), together with the TPR and FPR realised under automatic tracking (right, and section 4.4).

4.4 Segmentation of Ground Truth Plant Images

Before combining stages I and II of the algorithm and analysing overall performance, we test stage II on the *ground truth plant matter images* under automatic tracking by our Kalman filter algorithm [15]. We perform this experiment in order to assess stage II of the algorithm in such a way that is as far as possible independent of the image thresholding algorithm of stage I. The test is not entirely independent of the image processing errors, because they have an effect on the tracker's estimate of the crop grid position that is used in the feature clustering algorithm, but it does allow us to compare the true positive and false positive ratios for crop pixels directly with those found in the parameter selection experiments.

We use the Kalman filter's estimate of the crop grid position in conjunction with the size filtering and feature clustering algorithm of algorithm stage II. After this processing, we have two sets of classified pixels for each image sequence. The first set is C, the ground truth plant matter pixels that have been classified as

crop. The second set is W, the ground truth plant matter pixels that have been classified as weed. Given the two sets C and W, we can produce true positive (ground truth crop pixels that are classified C) and false positive (ground truth weed pixels classified as C) ratios for the automatic segmentation. These ratios are given in table 3 in the column marked 'automatic tracking'.

Before we compare the ratios from the tracking experiment with those from the parameter setting experiment, we reiterate the main differences between the two experiments. In the tracking experiment, the association region, within which all features are classified as crop, includes the uncertainty on the grid position, so will be larger than the corresponding region in the parameter setting experiment where the grid position was assumed to be known perfectly. We might expect that, as the association region expands, more image features will fall within it, so both TPR and FPR are likely to rise. The second difference is that the grid position in the tracking experiment is determined automatically by tracking the features derived from image processing, whilst the in the parameter setting experiment, the grid was placed by hand, and will be unaffected by any image processing errors.

If we now compare the tracking and parameter setting figures in table 3, we can see how these two experimental differences manifest themselves for each sequence:

Sequence A: The TPR drops and the FPR rises when the grid is tracked automatically. This sequence is the most difficult to track, because many crop plant features merge together so that feature centroids do not represent plant locations. Poor tracking is almost certainly the cause of the increased errors.
Sequence B: The TPR rises for the automatic tracker, where the association regions will be larger than in the parameter setting experiment owing to the increased uncertainty on plant position. The FPR is unaffected; this is a result of the low weed density in sequence B.
Sequence C: Both TPR and FPR increase under automatic tracking. This will be caused by the larger association region as it incorporates plant position uncertainty from the tracker.
Sequence D: As with sequence C, both TPR and FPR increase. Owing to the strong shadows present in this sequence, automatic tracking is difficult, so the uncertainty on individual plant position will be large; this is reflected in the dramatic rise in FPR.

The figures in table 3 show that the combination of size filtering and feature merging is very effective for classifying crop features, with true positive ratios in excess of 0.99 in for every sequence. The algorithm is less effective at weed pixel classification when tracking is difficult, as in sequences A and D, where the FPR rises to 15% and 20% respectively. This is not surprising because the success on the feature clustering algorithm hinges on the crop grid tracker providing good estimates of the crop position. However, when the tracking is easier, as in sequences B and C, the FPRs are much lower, 0.0% and 3.07% respectively.

4.5 Combining Stages I and II

The second segmentation experiment relies wholly on the thresholding and chain-coding algorithms and tests the full automatic segmentation algorithm that combines stages I and II. In the previous experiment, we knew that all the features presented for size filtering and clustering were true plant matter. In this experiment, some soil pixels will be misclassified as plant matter (and labelled C or W), and some plant matter pixels (crop or weed) will be labelled S. A suitable value for the threshold gain α for each sequence was determined from the slope of the ROC curves generated for each sequence [18] by using an empirical estimate of the Bayes costs and values and prior probabilities computed from the test data.

Each of the tables 4 – 7 presents the percentage of the ground truth crop, weed and soil pixels classified as C, W and S, together with the total number of ground truth pixels in each class from the ground truth images of sequences A – D. The numbers of pixels bordering ground truth crop and weed features are also given as an indication of the number of doubt pixels that have been ignored in the classification totals. Each image is composed of 384×288 pixels, although only pixels in the region of the image ($approx 65\%$) that will pass underneath the autonomous vehicle's treatment system (a bar of spray nozzles that runs along the front axis of the vehicle) are classified.

Perusal of the figures in tables 4 – 7 prompts a number of observations:

1. In every sequence, in excess of 98% of the soil pixels are correctly classified as S.
2. In each sequence, more crop pixels are misclassified as S than misclassified as W.
3. In each sequence, more weed pixels are misclassified as S than misclassified as C.
4. In sequences A and C, a greater percentage of crop pixels are correctly classified C than the percentage of weed pixels that are correctly classified as W.
5. In sequences B and D, a greater percentage of weed pixels are correctly classified W than the percentage of crop pixels that are classified C.
6. The number of doubt pixels that border ground truth weed features outnumber the total number of ground truth weed pixels in every test sequence.
7. The total number of ground truth crop pixels outnumber the doubt pixels that border the crop features in every test sequence.

Observations 1, 2 and 3 directly reflect the performance of the adaptive interpolated grey-level thresholding algorithm, which misclassifies a large percentage of the plant matter pixels as soil. This will obviously be the most common misclassification, because plant matter is most often seen against a background of soil rather than other plant matter. The observations do, however, highlight the fact that the plant matter/soil discrimination problem requires more attention if image segmentation is to be improved.

		Classified as			Number of	
		C (%)	W (%)	S (%)	pixels	border pixels
Ground truth	Crop	95.11	1.28	3.61	331,222	52,688
	Weed	10.50	51.88	37.62	505	1,373
	Soil	0.33	0.03	99.64	905,200	-

Table 4. Sequence A segmentation results, percentages of true numbers of crop, weed and soil pixels classified as C, W or S, and the number of pixels that border crop and weed features. There are 16 ground truth images for sequence A.

		Classified as			Number of	
		C (%)	W (%)	S (%)	pixels	border pixels
Ground truth	Crop	78.90	3.72	17.38	53,514	19,615
	Weed	0.0	81.8	18.2	934	3,455
	Soil	0.01	0.04	99.95	1,152,254	-

Table 5. Sequence B segmentation results, percentages of true numbers of crop, weed and soil pixels classified as C, W or S, and the number of pixels that border crop and weed features. There are 17 ground truth images for sequence B.

		Classified as			Number of	
		C (%)	W (%)	S (%)	pixels	border pixels
Ground truth	Crop	81.72	4.51	13.76	141,075	19,615
	Weed	3.93	56.90	39.17	17,160	18,544
	Soil	0.06	0.24	99.7	1,195,308	-

Table 6. Run C segmentation results, percentages of true numbers of crop, weed and soil pixels classified as C, W or S, and the number of pixels that border crop and weed features. There are 17 ground truth images for sequence C.

		Classified as			Number of	
		C (%)	W (%)	S (%)	pixels	border pixels
Ground truth	Crop	55.00	4.11	40.89	41,411	13,171
	Weed	6.31	73.52	20.17	1,046	2,202
	Soil	0.06	1.13	98.81	1,003,418	-

Table 7. Run D segmentation results, percentages of true numbers of crop, weed and soil pixels classified as C, W or S, and the number of pixels that border crop and weed features.There are 16 ground truth images for sequence D.

Observations 4 and 5 suggest that the larger plants seen in image sequences A and C are more easily identified than the smaller plants in sequences B and D. The reasons for this are unclear, but may be related to changes in the infra-red reflectance of the crop plants as they age.

Observations 6 and 7 show that the weed features, which are dominated by border pixels, are typically smaller than the crop features. This has already been illustrated in figure 2 and forms the basis of the size threshold algorithm.

If we ignore the crop and weed ground truth pixels that the segmentation algorithm labels S, we can construct true positive and false positive ratios for the crop and weed pixels that have been classified as plant matter (either C or W). These figures are given for each sequence in table 8 and show that those pixels which *are* identified as plant matter are separated into the crop and weed classes with some success. This allows us to conjecture that if plant matter/soil discrimination were more reliable then figures similar to those in table 3 might be obtained.

Sequence	TPR	FPR
A	0.9639	0.1683
B	0.9550	0.0
C	0.9477	0.0650
D	0.9305	0.0790

Table 8. TPR and FPR for the correctly identified plant matter pixels in sequences A–D.

5 Conclusions

We have used a novel two stage algorithm developed for a horticultural application to illustrate that breaking an algorithm down into its constituent components and testing these individually can provide a better understanding of overall behaviour. Analysis of the test results allows us to conclude that the majority of the errors in the system are propagated forward from stage I of the algorithm. It was seen that II performs effectively on the data that is correctly propagated form stage I, so algorithm development should focus on improving the plant matter/soil segmentation. Empirical discrepancy analysis based on ROC curves and type I and type II statistical errors was used for the individual binary classifiers, and overall tri-partite classification figures given for the full algorithm.

Acknowledgement

This work was funded by the BBSRC.

References

1. I E Abdou. Quantitative methods of edge detection. Technical Report USCIPI Report 830, University of California Image Processing Institute, July 1978.

2. D C Alexander and B F Buxton. Modelling of sinlge mode distributions of colour data using directional statistics. In *Proceedings Computer Vision and Pattern Recognition*, 1997.
3. Y Bar-Shalom and T Fortmann. *Tracking and Data Association*. Academic Press, New York, 1988.
4. H.R. Biller. Reduced input of herbicides by use of optoelectronic sensors. *Journal of Agricultural Engineering Research*, 71:357–362, 1998.
5. K Bowyer, C Kranenburg, and S Dougherty. Edge detector evaluation using empirical ROC curves. In *Proc. CVPR*, volume 1, 1999.
6. G J Edwards, C J Taylor, and T F Cootes. Improving identification performance by integrating evidence from sequences. In *Proc. CVPR*, volume 1, 1999.
7. H Freeman. On the encoding of arbitrary geometric configurations. *IEEE Trans. Elec. Computers*, EC-10:260–268, 1961.
8. D M Green and J A Swets. *Signal Detection Theory and Psychophysics*. John Wiley and Sons, 1966.
9. J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, April 1982.
10. R M Haralick. Performance characterization protocol in computer vision. In *DARPA Image Understanding Workshop*, 1994.
11. M G M Hunink, R G M Deslegte, and M F Hoogesteger. ROC analysis of the clinicl, CT and MRI diagnosis of orbital space-occupying lesions. *ORBIT*, 8(3), September 1989.
12. M Oren and S K Nayar. Generalization of the Lambertian model and implications for machine vision. *International Journal of Computer Vision*, 14:227–251, 1995.
13. B Rao. Data association methods for tracking systems. In A Blake and A Yuille, editors, *Active Vision*, chapter 6. MIT Press, 1992.
14. M J J Scott, M Niranjan, and R W Prager. Realisable classifiers: Improving operating performance on variable cost problems. In *Proceedings BMVC 1998*, volume I, pages 306–315, September 1998.
15. B Southall, B F Buxton, and J A Marchant. Controllability and observability: Tools for Kalman filter design. In M S Nixon, editor, *Proceedings 9^{th} British Machine Vision Conference*, volume 1, pages 164–173, September 1998.
16. B Southall, T Hague, J A Marchant, and B F Buxton. Vision-aided outdoor navigation of an autonomous horticultural vehicle. In Henrik I Christensen, editor, *Proceedings 1^{st} International Conference on Vision Systems*. Springer Verlag, January 1999.
17. B Southall, J A Marchant, T Hague, and B F Buxton. Model based tracking for navigation and segmentation. In H Burkhardt and B Neumann, editors, *Proceedings 5^{th} European Conference on Computer Vision*, volume 1, pages 797–811, June 1998.
18. J B Southall. *The design and evaluation of computer vision algorithms for the control of an autonomous horticultural vehicle*. PhD thesis, University of London, 2000.
19. M R Spiegel. *Probability and Statistics*. Schaum's Outline Series. McGraw Hill, 1980.
20. B van Branniken, M Stavridi, and J J Koenderink. Diffuse and specular reflectance from rough surfaces. *Applied Optics*, 37(1):130–139, January 1998.
21. H L van Trees. *Detection, Estimation and Modulation Theory, Part I*. John Wiley and Sons, 1968.
22. Y J Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.