RESEARCH ARTICLE OPEN ACCESS

Genetic Algorithm Tuned SVM Classifier for Weed Species Recognition

W.K Wong, Ali Chekima, Muralindran Mariappan,
Brendan Khoo, Manimehala Nadarajan

Faculty of Engineering
Universiti Malaysia Sabah
Kota Kinabalu,
Sabah-Malaysia

ABSTRACT

Weed detection using image processing is a progressive research area which can revolutionize crop production and increase efficiency in herbicide usage. A SVM classifier was applied in our research to classify and to detect the weed type for weed scouting and spot weeding purposes. However, parameter fine tuning and feature selection for SVM is a complex procedure that is still an active area of research. A wrapper type feature selection and parameter fine tuning were proposed for the SVM training using Genetic algorithm to reduce the features and consequently over fitting of classification. The resulting classifier requires 2 features out of the 68 features. The performance of the classifier were optimal with additional 200 test samples with 96% correctly classified for positive label and 100% for negative label.

Keywords:- Support Vector Machine (SVM), weed scouting, weed species recognition, Genetic Algorithm

I. INTRODUCTION AND RELATED WORKS

The reduction of herbicide without affecting the crop yield is a promising area of research which can be achieved using image processing technology. Various on-going research work have been proposed using textural and shape analysis. Most research work have focused on distinguishing between Crop – weed [1][2] and distinguishing monocotyledon – Dicotyledon weeds [3][4].Currently several researchers have provided a fairly accurate recognition to recognise between the classes of weeds.

The areas of applications of weed recognition may differ in the types of weeds to be distinguished, maturity/density of weeds and the image acquisition device setup. For mature weeds patches, texture analysis [1][2][3]were often used to distinguish between weed categories. For recognition of weeds in early stages of post emergence, shape analysis are often used. The shape analysis such as regional shape features [4][5], image moments and fractal dimensions [6] were used in respective researcher work. Neural networks and SVM were among classifier types used to distinguish between the classes of weeds such in [2][3]. In this aspect, neural network has more advantage as it can implement multiclass classification while SVM is generally a binary class classifier. However, to recognize various types of weeds SVM can be ensemble using 'One against All' (OAA) strategy or 'One against One' (OAO) strategy. In order to implement these strategies with optimal results, individual SVM trained have to display a high recognition rate as the overall ensemble of SVM classification can be effected by individual SVM performance.

II. RESEARCH BACKGROUND.

Various features have been tested to recognise between individual species and classes of weeds. Among features tested were fractal dimensions, image moments, Elliptical fourier descriptors and various region based shape descriptors. In our previous research [7] various features were combined to recognize the individual weed species. The SVM were trained and optimised using Genetic Algorithm to select the features and to fine tune the soft hand over constant, C and RBF kernel parameter, σ. However, these weeds have clear distinctive shape enabling easy recognition provided the test samples and training samples have low variances. Further reduction of features were not performed in above mentioned research work. Although these features produced high recognition rates, further analysis can be performed to eliminate non – discriminant features.

This paper is an extension of the existing work to train the SVM using similar Genetic Algorithm methods to recognise *Amaranthus Palmeri*, Monocotyledon weeds and *Agerantum Conyzoides* among other type of post emergence weeds. Both Ageratum Conyzoides and Amaranthus Palmeri are known to be highly invasive species. Although selective spot weeding technology normally requires only distinguishing between broad leaf and narrow leaf weed species, additional information on the types of weeds can enable the farmers to plan out various strategies to eliminate the weeds that are found to be resistive to the herbicide applied. To optimise the overall classifier, each individual classifier needs to be optimised. Hence, this paper focus is to test the Genetic algorithm with varying weightage to optimise the recognition and to reduce non discriminant features. This research work will focus on optimising the SVM to distinguish between monocotyledon – dicotyledon weed using genetic algorithm.

III. METHODOLOGY

Samples were collected from various crop field. Among the species in the training, verification and testing sample images are the *Amaranthus Palmeri* , *Agerantum Conyzoides* . among other weed species such as *Phylanthus Urinuria* and other dicotyledon and monocotyledon weeds. The weed seedling samples were sampled from crop farms which were collected and transplanted into containers for image acquisition. The samples are collected after emergence level. The camera used is the Logitech c615 Web cam with 1920 x 1050 pixels and located 28 centimetres from the soil. The vegetation were segmented from the soil by using excessive green index as shown in eqn 1 as applied by Meyer *et al.*,in [8] in vegetation segmentation from the soil. Other vegetation indices and segmentation were discussed in and compared in [9]

$$Ex\text{-}G = 2G\text{-}R\text{-}B \qquad \textbf{eqn}\ (1)$$

Where R,G,B are normalized pixel value of red, green and blue pixels where $R = \frac{r}{r+g\_b}\ G = \frac{g}{r+g\_b}\ ,\ B = \frac{b}{r+g\_b}$ *and r,g,b are the non-normalized values of RGB pixels.*

$$fn\big(x(i,j)\big) = \begin{cases} 1, if\ x(i,j) \gg 0.1 \\ 0, if\ x(i,j) < 0.1 \end{cases} \qquad \textbf{eqn}\ (2)$$

A total of 68 features were included for feature selection which were the rotation/scale invariant shapes, fractal dimensions, HU's moments, elliptical fourier coefficients (1-10th coefficient) and skeleton statistics (mean, variance, skewness, maximum distance and skeleton area) distance to centroid and colour features were considered for weed overall shape and leaf shape analysis.

A total of 8 region based shape features were proposed for the classification as shown from eqn. 16 to eqn.21. The shape features were acquired by binarizing the images from the excessive green filtered images using threshold methods.

$$\text{elongation} = \frac{width\ of\ boundary\ box}{length\ of\ boundary\ box} \qquad \textbf{eqn}\ (3)$$

$$\text{Solidity} = \frac{filled\ area}{convex\ area} \qquad \textbf{eqn}\ (4)$$

$$\text{Eccentricity} = \frac{\text{distance between the foci}}{major\ axis\ length} \qquad \textbf{eqn}\ (5)$$

$$\text{extend} = \frac{area}{total\ area\ of\ bounding\ box} \qquad \textbf{eqn}\ (6)$$

$$\text{compactness} = \frac{perimeter}{area} \qquad \textbf{eqn}\ (7)$$

$$\text{squareness} = \frac{convex\ \ area}{bonding\ box\ size} \qquad \textbf{eqn}\ (8)$$

$$\text{Convexity} = \frac{perimeter}{convex\ perimeter} \qquad \textbf{eqn}\ (9)$$

$$\text{circularity}= \frac{\text{filled area}}{\text{equivalent diameter} * pi * 4} \qquad \textbf{eqn } (10)$$

where boundary box is the smallest rectangle that contains the binarized image, filled area is the amount of pixels in the binarized image,convex area is the polygon containing the binarized image. equivalent diameter is the diameter of the smallest circle containing the binarized image.

Fractal dimensions are parameters that can be applied as shape descriptors in both 2-dimensional and 3-dimensional shape .Various works have feature the application of fractal dimensions for weed classification such as in [10]. A known way of calculating fractal dimensions is using the box counting methods. Where N is denoted as the number as size δ squares required tofill the specified shape, d, fractal dimensions is as defined in eqn 11.

$$d = \frac{\log N(\delta)}{\log(\delta)} \qquad \textbf{eqn } (11)$$

The HU s moments invariants, and Elliptical Fourier coefficients were explained in [11] and respectively and were included in the feature vector for training. The 5 skeleton statistics are mean distance between the skeleton line and boundary, maximum distance between skeleton and boundary point and size of generated skeleton, variance of distance from boundary to skeleton and skewness of the skeleton statistics. Similar skeleton based statistic features were applied in [12] producing reliable results thus further justifying the application of this feature.
The centroid to boundary features as indicate in the table 1 where the boundary to centroid features. The 2 features in this category are the mean values from boundary to centroid and the max value. The colour features are the mean values of the normalized R,Y,B and saturation value of pixels within the binary images. As shown in the Table 1, this features are only used in the overall weed shape analysis and not the segmented leaf analysis.

Support vector machine (SVM) classification applied for this classification classifies by constructing a hyperplane / hyperplanes in high dimensional spaces of data. The SVM classifier finds a marginal line that defines the space between the two classes which is known as margin. The points on the margin are known as support vectors. A best hyper plane is the hyper plane that represents the largest separation between the 2 classes of data. A larger margin space would enable higher classification rate. However, this would compromise on the misclassification of points. A trade off on this can be achieved by fine tuning the *C*, parameter. SVM can be fine tune by setting the soft hand over coefficient, *C* and the alpha value (for Radial basis function only). Another value that can be fine tune was the σ value, which only applies if RBF kernel is used in the SVM.

Genetic algorithms are loosely based on the evolutionary process of nature in selecting the fittest genes from the gene pool and consequently changing the species features. This 'Darwanian' concept is utilized by encoding chromosomes as seven chromosome to represent the features, alpha value of RBF kernel, and the soft handover coefficient of the SVM, *C.).* To obtain optimal recognition in each successive cross over, a chromosome is mutated by randomly combining the binary variable of two parent chromosome. The chromosome description is shown in table 2. The SVM is trained with 200 features rows and verified with 200 feature rows. Both the training data and the verification data are balanced (50%-50%). The optimised SVM are tested externally with external test sets (200 data sets) in which 100 data sets belong to the *specified* weeds (positive label) and 100 from other weeds (negative label).

The random selection of features and values were repeated until the average values change of the population fitness is less than 0.5 %. GA algorithm randomly select and unselect the features by turning the bits 'on' and 'off' on the binary chromosome. The first 16 bits represents the shape features /fractal dimensions/Hu moments. The 2nd, 3rd and 4th chromosome represents the elliptical Fourier (40 bits). The 5th chromosome represents the pixel average values of skeleton statistics. The 6th and 7th chromosome values are integer's value of the σ value and *C* (soft hand over constant). The fitness function is the classification rates (%) for the verification data sets. The chromosome detail containing the features is shown in Table 1.

Table 1: Bit description of features for dilated binarized overall weed shape (left) and binarized weed leaf (right)

Chromosome no.	Type	Bit length	Description of representation
1	Binary	17	Shape (9bits) features/fractal (1 bit)/ Hu moments (7 bit)
2	Binary	15	Elliptical fourier descriptors (40bits)
3	Binary	15	
4	Binary	10	
5	Binary	7	Skeleton statistics(5 bits) Boundary to centroid (2 bits) Number of set pixels(1 bit) Colour statistics(4 bits)
6	Integer	n/a	σ value
7	Integer	n/a	C value

The SVM is trained with 100 data vectors and further verified with another 100 data vectors. The classification rate of the verification vectors were calculated to determine the fitness function as expressed in equation 13. WA is the weightage coefficient introduced to trade of between the number of features and recognition rates of verification sets. The WA coefficient was changed from 1.0 to 0.3 with an interval of 0.1. The reduction in WA coefficient will contribute to the weightage of fitness function to reduce feature number and eliminate non discriminant features. The increment of WA will have a vice versa effect on the fitness function as more weightage will be given to the increment of recognition rate of verification sets. By adjusting and varying the WA coefficient, various combination of feature were selected as shown in table 2.

$$fitness\ function = WA\ x\ (recognition\ rates) + (1 - WA)\ x\ \left(\frac{no\ \ of\ features}{no\ of\ features - no\ of\ selected\ features}\right) \qquad \textbf{eqn}\ \ (13)$$

Table 2: Features selected from Genetic algorithm

	No of features selected									parameters	
W_A	Regional shape features	Fractal	Hu Moment Invariants	Elliptical Fourier descriptors	Skeleton stats	Boundary to centroid	Number of set pixels	Colour stats	Total Feature	σ	C
1.0	2	0	0	25	1	0	0	0	**28**	**1.024**	**0.538**
0.9	5	1	0	14	0	1	0	2	**23**	**1.024**	**0.765**
0.8	1	5	0	13	3	0	0	1	**23**	**1.023**	**0.962**
0.7	4	5	0	16	1	0	1	0	**27**	**1.021**	**0.208**
0.6	1	1	0	12	1	0	0	0	**15**	**1.020**	**0.976**
0.5	1	0	0	11	0	0	0	0	**12**	**1.024**	**0.010**
0.4	3	7	1	6	1	0	0	0	**18**	**1.024**	**0.012**
0.3	1	1	0	0	0	0	0	0	**2**	**1.024**	**0.010**

The GA optimised SVM to recognise the individual weed species is has enabled an individual weed classification with 100% classification by selection of the features for training and the tuning of the soft handover value- C and the σ value of the RBF kernel applied on the SVM. It is noteworthy that GA does not always give the 'global best' answer as they tend to converge on the 'local best' configuration. Hence, the GA was run for several times until the best result was achieved. The weightage coefficient , WA was changed from 1.0 to 0.3 and with each with each weightage coefficient value, the GA was applied for feature selection and optimisation. The resulting SVM classifier is further tested with additional testing sets. Table 3 shows the testing with additional testing sets. Positive false and negative false refers to samples that .The positive false are due to the samples that are almost similar shapes as the *Amarathus Palmer sp.* leaves. The negative false are due to the high variances in the leaf shapes.

A probabilistic SVM output was implemented using Platt's proposed posterior output in [13]. The classifier output is set to positive label and negative label at a threshold value of 0.5 as expressed in eqn. 14 during testing phase with external samples after the SVM development.

$$\text{fn}_{\text{prob}} \begin{cases} 1 \text{ if prob } \geq 0.5 \\ 0 \text{ if prob } < 0.5 \end{cases} \qquad \textbf{eqn (14)}$$

IV. RESULTS

The WA is reduced at an interval of 0.1 from 1.0 to 0.3. At value lower than 0.3, it was found that the GA converge to deselect all features (no features selected) hence causing an error in feature selection and further generating a fitness value. Figure 1 shows the effects of adjusting the WA weightage coefficient. Figure 1 (left) shows the increment of recognition rate of verification sets as the WA value is decreased. It was also observed in figure 1(right) that the number of features selected decreases as the value of WA was decreased from 1.0 to 0.3. At WA=0.3, only 2 features were selected. This shows that there is a reduction of non – discriminant features by decreasing the WA value. By decreasing the number of non-discriminant features, the recognition rates were also increased thereby increasing the overall fitness function. The two feature selected at WA= 0.3 were the solidity and first HU moments.

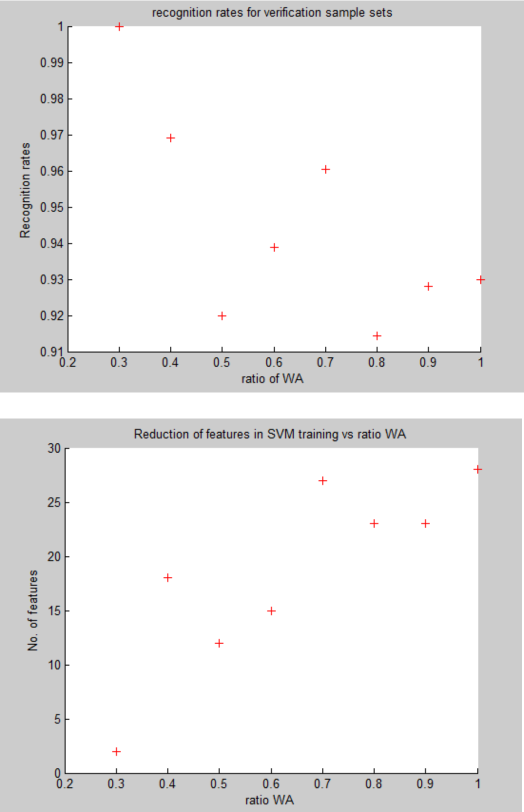

Figure 1: (First) varying recognition rates of verification sample sets with varying weightage ratio, WA (Second) varying number of features selected with varying weightage ratio, WA

The trained and optimised SVM were further tested with 200 external sample set. The test images were crop images of various weed species. Table 3 shows the recognition rates of SVM trained with GA at various WA values.(from 1.0 to 0.3 with interval value of 0.1). As shown in table 3, the SVM trained with WA= 0.6 has the lowest mean error for positive label followed by the SVM trained with WA= 0.7, 0.8, 0.9. The positive label mean error for SVM trained with WA=0.3 was generally higher than other SVM configurations. From table 3, the negative label was found be decreasing with the reduction of WA value. Further observation showed that there is a correlation between the reduction of features and reduction of mean error for negative samples although the mean error for positive samples did not correlate with the reduction of

Feature. The results in table 3 also showed that the SVM trained with WA=0.3 has the highest classification rate for negative label samples although it did not reach 100% classification for positive label samples sets.

Table 3: a) Confusion matrix of GA feature selected SVM testing with varying weightage, W_A coefficient

W_A	(+) true %	(-) false %	(+) false %	(-) true %	No of features	Mean error (positive)	Mean error (Negative)
1.00	100	0	33	67	28	0.2203	0.3084
0.9	100	0	35	65	23	0.2247	0.3735
0.8	100	0	52	48	23	0.1925	0.3300
0.7	100	0	67	33	27	0.1828	0.3974
0.6	100	0	40	60	15	0.1746	0.2706
0.5	90	10	0	100	12	0.2948	0.2053
0.4	100	0	10	90	18	0.3624	0.2835
0.3	96	4	0	100	2	0.2720	0.1964

The results show that the feature reduction can be enhanced by tuning the weightage coefficient, WA. The non – discriminant features present has contributed to over fitting thereby increasing misclassification during test phase. The elimination of these features, the effect of over fitting has been reduced as demonstrated by the decrement of misclassification rates with external sample sets. The best SVM configuration was achieved at when SVM was trained and optimised with WA=0.3.

V. CONCLUSION

The paper proposed and tested a GA feature selection and parameter fine tuning for weed detection using SVM. Although SVM are known to be robust and resistant to over fitting, it still requires a careful selection of C and σ parameters. In this paper, a simultaneous feature selection and fine tuning of SVM parameters were performed using Genetic Algorithm. From the results shown, it was found that the optimised SVM classifies with 2 features which were the solidity feature and first HU moments. The optimised and trained SVM with various configurations were tested with additional sample sets. The results show that the SVM with the least features performed the best from the classification rates. In conclusion, the solidity and first moments were shown to be the best features to discriminate between the Monocotyledon and Dicotyledon weeds. More work can be done using the similar methods to train the other SVM to distinguish other weeds species.

REFERENCES

[1] Kianni S, 2012, Crop-Weed Discrimination Via Wavelet-Based Texture Analysis, *International Journal of Natural and Engineering Sciences vol.6 (2). pp: 7-11 , 2012*

[2] Wu . Lahlan and Wen. Youxian, 2009, “Weed/Corn Seedling Recognition By Support Vector Machine Using Texture Features”, African Journal of Agricultural Research vol. 4(9) pp. 840-846

[3] Tang.Lie, Tian. Lei F, Steward. Brian L, 2003, ‘Classification of Broadleaf And Graff Weeds Using Gabor Wavelets And An Artificial Neural Network’, American Society of Agricultural Engineers. VOL 46(4) . pp.1247-1254.

[4] Kamal N. Agrawal, Karan Singh, Ganesh C. Bora and Dongqing Lin,,2012,Weed Recognition Using Image-Processing Technique Based on Leaf Parameters. Journal of Agricultural Science and Technology,Vol. B 2 (2012) , pp:899-908

[5] Kianni.S and Jafari. A, 2012, ‘Crop Detection and Positioning in the Field Using Discriminant Analysis and Neural Networks Based on Shape Features’. Journal of agricultural science technology. vol 14. pp:755-765

[6] Lanlan Wu, Youxian Wen, Xiaoyan Deng2 and Hui Peng,2009,’Identification of weed/corn using BP network based on wavelet features and fractal dimension’, Scientific Research and Essay Vol.4 (11), pp. 1194-1200,

[7] Wong W.K, Ali Chekima, Muralindran Mariappan, Brendan Khoo, Choo C.W, Manimehala Nadarajan, 2014, Genetic Algorithm optimization and feature selection for a support Vector Machine weed recognition system at critical Stage of Development .World Applied Sciences Journal 30(12).pp:1953-1959

[8] Isabelle Phille, Thomas RATH , 2002, ‘Improving discrimination in image processing by use of different colour space transformation. Computer and electronics in Agriculture.(35) pp: 1-15

[9] Meyer.G.E, Metha.T, Kocher M.F, Mortesen D.A, Samal A, 1998, ‘Textural Imaging and Discriminant analysis for distinguishing weeds for spot spraying. ‘ASAE, st. Joseph V. 41, pp 1189-1197

[10] Solahudin .Mohamad , I Wayan Astika, Kudang Boro Seminar, Agus Buono,2010,’Weeds and Plants Recognition using Fuzzy Clustering and Fractal Dimension Methods for Automatic Weed Control’, AFITA 2010 International Conference, pp :110-112

[11] HU.M, 1962, Visual pattern recognition by moment invariants. IRE trans. Inf. Theor. IT-8:179187

[12] [12] Till Rumpf , Christoph Römer , Martin Weis , Markus Sökefeld , Roland Gerhards , Lutz Plümer 2012, Sequential support vector machine classification for small-grain weed species discrimination with special regard to Cirsium arvense and Galium aparine. Computers and Electronics in Agriculture (80)pp:89-96[13] Platt. John c.,1999,Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods (1999). ADVANCES IN LARGE MARGIN CLASSIFIERSpp:1-11