TEXTURE AND SHAPE ATTRIBUTE SELECTION FOR PLANT DISEASE MONITORING IN A MOBILE CLOUD-BASED ENVIRONMENT

Punnarai Siricharoen, Bryan Scotney, Philip Morrow and Gerard Parr

School of Computing and Information Engineering
Ulster University, Coleraine, UK

ABSTRACT

We focus on feature extraction and selection to best represent texture and shape properties of plant diseases in an image-based leaf monitoring system implemented in a mobile-cloud environment. A number of textural and region-based features are aggregated from previous studies; also we introduce mean and peak indices of histogram-of-shape as disease property representations along with the proposed and enhanced shape features based on diseased regions. A total of 260 colour-based attributes and 163 shape attributes are searched to find the best potential features based on different aspects: probability of feature error, correlation, targeted-class relevancy and the separability quality of a feature. Experimental results show that the best selected feature set which combines colour-based and shape features yields high classification accuracy on wheat disease images captured by a smartphone camera and also provides insights into potential sets of features to be further implemented as a lightweight standalone mobile application.

Index Terms— histogram of shape features, textural features, feature selection, pathological plant monitoring

1. INTRODUCTION

Plants play a fundamental role in supporting all kinds of life on the planet, particularly through food and medicine. Plant diseases affect human society in terms of economic loss and public health, including insufficient nutrition. Closely and continuously monitoring plant health is required to prevent such damage. For many diseases, plant pathologist familiar with the appearance and characteristics of unhealthy plants are required to identify diseases [1]. However, this is labour-intensive and cost-ineffective, especially in large-scale agricultural business or in remote areas. Also, disease detection in its late stages might indicate severe and irreparable loss. When crops are infected, characteristic symptoms are visible which differ from healthy crops in recognizable ways. Thus, an automated detection imaging system can be a useful aid to local farmers for early detection. Mobile capture devices are recently ubiquitous and affordable, and cloud computing technologies have been rapidly developed. This work leverages imaging techniques to integrate with mobile image capture and mobile cloud computing, resulting in an accurate plant health monitoring application.

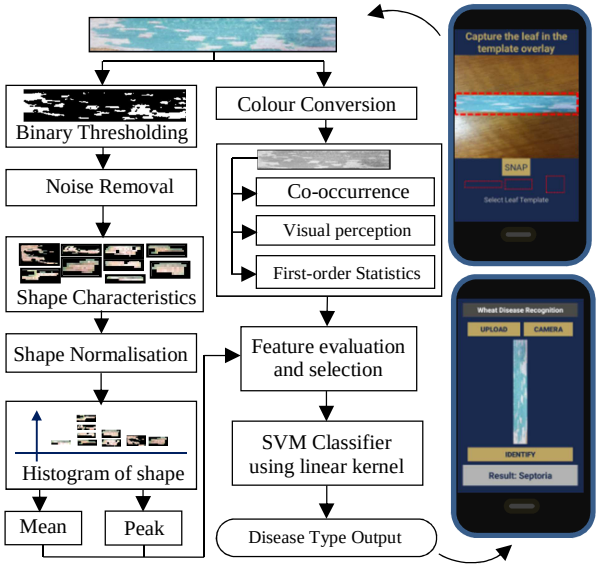


Fig. 1. Plant Disease Monitoring System based on colour-based features and shape characteristics of disease patterns

The challenges of crop disease identification using imaging techniques include the variety of plants and diseases, the high degree of similarity between different diseases, the differences in grading of a disease, and the variety of image capture conditions. Previous research studies [2]–[5] have developed an automated system through data acquisition under constrained conditions to accurately extract textural information from the diseased image. The potential sets of features resulted in combinations of basic first-order statistical features, features derived from co-occurrence matrices and shape features to describe different types of diseases. Although the combinations offered accurate classification, individual sub-features contribute to the system differently. Tain et al. [4] empirically selected some sub-features from the combination, whereas, Sarayloo et al. [5] deployed minimal-redundancy-maximal-relevance to evaluate sub-features and rank them before selecting the best features to represent the diseases.

We propose an automated disease recognition system as summarised in Figure 1. The cloud-based classification system uses the previously studied sub-features, including textural features derived from a co-occurrence matrix, and first-order statistical features, including visual perception features. Also, we introduce peak and mean representation of a histogram of shape regarding the natural distribution of diseases. Finally, the

contributions of sub-features are evaluated across different aspects, including individual classification performance, separability quality, and redundancy-relevance. For feature evaluation methods, we use minimal-redundancy-maximal-relevance (mRMR) [6], ReliefF [7] and probability of error combined with feature correlation [8]. Then sequential forward selection is employed to retain only a small set of the best features. These flexible feature selection techniques maintain the original features, so we can still select these potential features as the lightweight features for a further mobile-standalone application. For the current cloud-based application, an image of a leaf is captured using an overlay template (we can alter the leaf template in figure 1), and then delivered over a network via HTTP. Only the selected potential features are calculated and passed through an SVM classifier before the output result is returned to the smartphone. To measure the performance of the system, we have experimented with real-world datasets of wheat diseases captured by a standard smartphone and acquired by the UK Food and Environmental Research Agency (FERA) [9]. The final selected subsets of textural features and shape are combined to demonstrate high classification accuracy from the system experimentation. Additionally, when these sets are employed in the mobile-cloud computing environment, experimentation on different mobile capture devices yields promising outcomes.

2. FEATURE GENERATION

Many features derived from co-occurrence matrices, colour and shape have been shown in previous research [2]–[5] to be robust representational attributes for plant diseases. In this section we discuss aggregation of a range of features, which are categorized into two groups: colour-based features and region-based features. We also introduce peak and mean indices of histograms of region-based features to represent shape characteristics for each leaf and propose two types of modified principal axis ratios.

2.1. Colour-based features

Textural features and statistical attributes rely on selected colour components. The relevant features in this work include 13 textural features, 4 first-order statistical attributes and 3 visual perception features which are detailed below.

Textural features are developed through a grey-level co-occurrence matrix which computes frequencies of two pixels with quantized intensity levels i and j and separated by distance d and orientation θ [10]. These features include, but are not limited to, homogeneity, correlation, contrast, energy, etc.

First-order statistical features (or colour features) measure basic information of colour distribution, but are powerful properties in many applications. Mean, standard deviation, skewness and kurtosis are considered in our system.

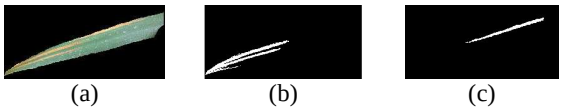
Visual perception features: Tamura [11] proposed 6 computational features including coarseness, directionality, contrast, line-likeness, regularity and roughness. The three first features are considered in the system; while the latter three which are derived from the three former features are neglected.

2.2. Region-based features

To segment the disease region, Otsu's binary thresholding is applied on the Cb and Cr colour components (of the YCbCr colour model) which are shown to be robust to different lighting condition [12]. Figure 3 illustrates segmented disease samples of three types of wheat leaves, where the individual disease patch (spot) is computed for 15 different shape features.

Shape features Fifteen different shape features are used to represent diseases (1) principal axis ratio is a ratio of major axis length (L_m) and minor axis length (L_n) of a disease patch. (2) Area ratio is a ratio of disease area (A_D) by leaf area (A_L). (3) Circularity is the distance between foci of the ellipse shape (D_F) divided by major axis length (L_m). (4) Compactness measures a ratio of A_D and convex area (A_C) of the diseases. Since a captured leaf is in a template overlay, the leaves will have consistent orientation. (5) Orientation is measured as the direction of the particular disease patch. (6) Complexity measures the ratio of square of the disease perimeter (P_D^2) and disease area (A_D). (7) Equivalent diameter is also calculated to specify the radius of a circle that could have the same area as the disease spot. (8) Hydraulic radius computes the ratio of A_D and P_D. (9-15) Seven Hu's moment invariants [13] of geometric shape were demonstrated to be independent to a figure's size, orientation, or position.

Two more region-based features, modified principal axis ratios, are introduced based on the distributed nature of disease structure. The first modified principal axis ratio (PAR+AR) is weighted by disease area; the larger the disease patch, the higher the impact on the histogram of the feature. Although morphological techniques are applied to the segmented diseases they still are unable to be segmented accurately as shown in Figure 2(b) segmented from 2(a). Thus, another modified principal axis ratio is weighted by disease area and the degree of solidity of the disease patch (PAR+AR+SLD); the more compact the disease; the higher the impact on the histogram.



(a) (b) (c)

Fig 2. Two main disease patches (b) and (c) are segmented from (a) where (c) is the expected segmented result but (b) is not segmented properly.

Histogram of Shape Disease segmentation in Figure 3 (middle row) results in more than one of the disease patches including noise. Noise removal is performed by eliminating the patch which has an area less than {Area Cut factor (AC) x Largest Patch Area}. Then, a histogram of shape properties is constructed from the shape features of the remaining disease patches. Instead of using histogram values as features, we introduce peak and mean indices to represent the histogram. The histogram of principal axis ratio of 6 wheat leaves is shown in Figure 3 (bottom row). The red arrows point at the peak index of each histogram. It can be roughly concluded that the ratio of the yellow rust leaves have the highest peak index value, whereas Septoria diseased patches has lower peak index and non-diseased patches containing mostly noise have the lowest peak ratio index.

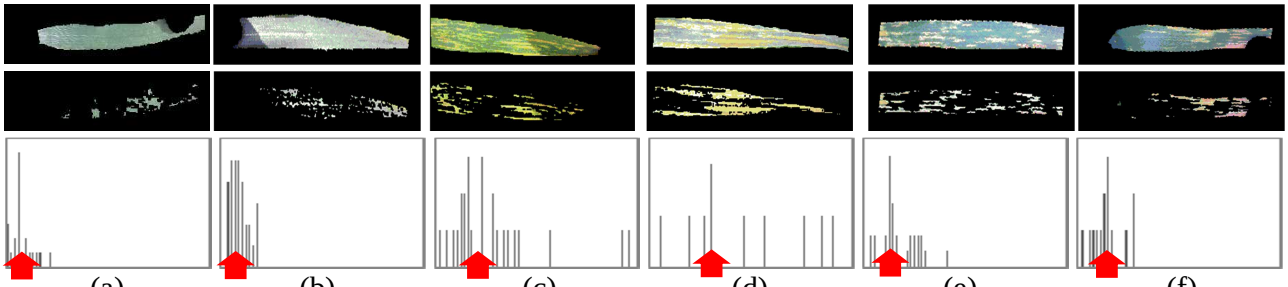


Fig. 3 Examples of healthy green leaves (a) and (b), yellow rust diseased leaves (c) and (d), and Septoria diseased leaves (e) and (f) (Top row). The middle row shows the disease region segmentation results corresponding to the leaves in the top row. The histogram of principal axis ratio is shown in the bottom row with the the red arrows pointing to the peak index of the histograms.

3. FEATURE SELECTION

Prior to feature selection feature evaluation is applied to rank the main twenty colour-based features and fifteen region-based features. Then, the ranked features are selected sequentially using a forward selection technique to eliminate the less important features.

Feature Evaluation: Three methods are considered in the system with the aim of improving the overall classification accuracy rate. (1) The weighted sum of Probability of Error rate (POE) and average correlation coefficient (ACC) are applied to assess each feature. POE is first used to select the first feature and the following feature on the rank is based on 90% of POE and 10% of ACC measures; the weight values were demonstrated in [8]. (2) ReliefF measures 'distinguishability' in a feature by considering the differences of the current instance with the K nearby instances from the same class and another K nearby instances from different class. Although ReliefF exploits the information locally, the combined context will provide a global view of the information [7]. (3) Max-relavance-min-redundancy (mRMR) considers the degree of relevance between a feature and a targeted class and also the redundancy between features based on mutual information [6]. The difference in the two aspects of relevance and redundancy is maximized for the top ranked features from mRMR assessment. From a complexity aspect, although POE is calculated from an individual feature, the POE+ACC feature assessment method has higher computation cost compared to others as it requires learning and testing of a classifier to calculate an error rate of a feature.

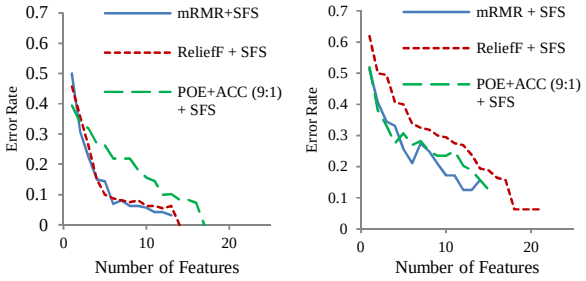
Feature Selection: The selected number of feature to be used in training a classifier is critical. After features are ranked, the sequential forward selection (SFS) method is applied to select features from the rank sequentially that meet the criteria. The first feature in the rank is an initial set of SFS to be evaluated and the next selected feature is the next feature in the rank that when combined with the first one improves the classification rate. The selection search technique has lower time complexity (O(N)) compared to exhaustive search.

4. EXPERIEMENTATION RESULTS

The performance of our proposed system is evaluated at the server-side of the system environment. The wheat images in our experimentation were captured in a wheat field using a standard smartphone by a FERA researcher [14]. This dataset comprises 160 labelled images (816x612 pixels) of which 50 are healthy green leaves (GL), 55 display Septoria (ST) disease and 55 display yellow rust (YR) disease. The primary leaves or main leaf in each image is analysed in our system as labelled by Gibson et al. in [14]. The testing scheme is 5-fold cross-validation using SVM classifier (with a linear kernel). The same testing scheme is employed in POE+ACC feature evaluation and sequential forward feature selection.

Two sets of features to be selected include colour-based features and shape features. Thirteen co-occurrence, four statistical and three visual perception features are transformed into thirteen colour components (Grey, RGB, YCbCr, L*a*b*, and HSV), giving a total of 260 ((13+4+3)x13) colour-based features to be selected. There are fifteen diseased-shape features including two modified principal axis ratio varying four different noise removal factors (AC = 0, 0.01, 0.05, 0.1) which are represented by peak index or mean value, thus 136 (17x4x2) region-based features are to be selected in the system. A bin size of 50 for the shape histogram is selected. ReliefF feature evaluation is based on K=5 neighbours to be considered.

Figure 4 shows the average testing error rate at a certain number of features from 5-fold cross-validation. Only less than 20 features out of 260 colour-texture features are enough to represent the diseases and at some folds it meets the lowest 0% error rate by ReliefF and POE+ACC feature selection. As shape features rely on disease segmentation, the lowest error rate is only less than 10% using ReliefF. The number of features required to meet the lowest error rate for shape features is around 20 features out of 136. Considering the performance of each selection scheme, mRMR+SFS reached its lowest error at less required number of features; for this dataset, ReliefF+SFS selection provides the least error rate at some folds. The results show that the combination of shape features and colour-based features increase overall accuracy. However, the best combined features are selected by mRMR+SFS feature selection as shown by the ROC curves in Figure 5; AUC values for healthy green leaf (GL) shows 100%, and 99.31% for Yellow rust disease and 98.79% for Septoria disease. The selected features improved the performance given in [14]; however, the primary leaves we considered are perfectly segmented.



(a) (b)

Fig 4. Comparison of average error rate at selected number of features (a) from 260 colour-based features from different feature selection methods (b) from 163 shape-based features from different feature selection methods

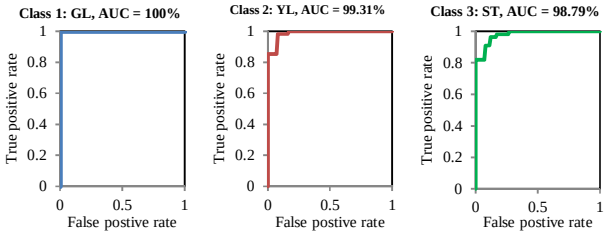
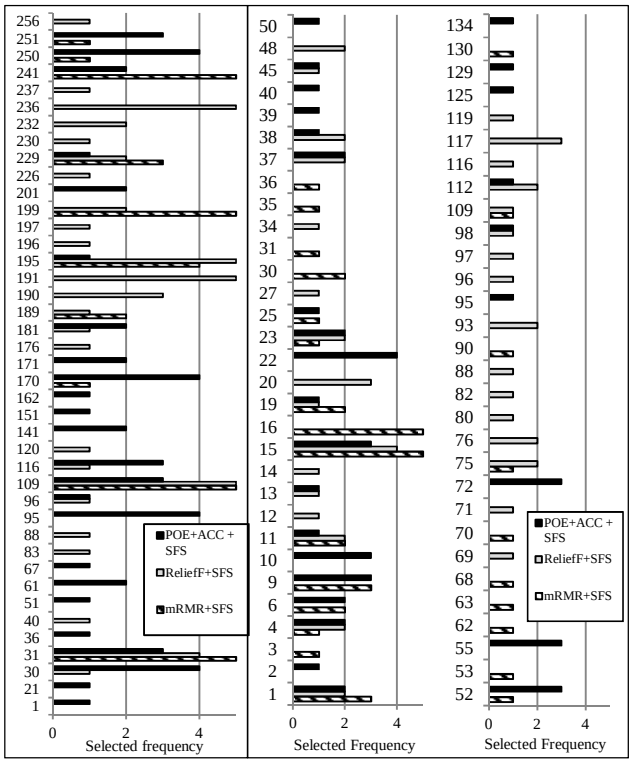


Fig 5. ROC curves of the combined colour-based features and shape features from mRMR + SFS feature selection method

Figure 6 displays the top-10 selected features from 5-fold scheme selection (~50 selected features per selection method). For colour-based features in Fig. 6(a), generally, chromatic components (R, H, Cb, a) are selected more than intensity components (Y, Grey, V). Homogeneity and entropy-based textural features are frequently presented (#31,#109,#191, #229,#241,#251). Other potential features include standard deviation, and contrast (#195) from Tamura's features (#199). For shape features in 6(b), hydraulic radius (#15,#30,#45,#75), principal axis ratio (#1,#16,#46) and some Hu's moments are selected by three evaluation aspects with high frequencies. The mean representations are selected more than peak index especially at #1-#15. No noise removal (AC = 0) is the most chosen in general. The higher AC factor, the less features are selected. Mean of PAR+AR+SLD distribution is chosen once based on error rate and correlation (POE+ACC).

5. CONCLUSION

The analysis of plant disease recognition system shows that the combined colour-based and shape features are powerful feature sets to represent plant disease patterns and achieve the high classification accuracy. These potential colour-based features includes homogeneity and entropy-based features (co-occurrence matrix), standard deviation of colour components and visual perception contrast features which are amongst commonly selected features regarding feature quality, target relevancy and accuracy aspects. The best selected shape features include hydraulic radius and principal axis ratio which describe how complicated and how elongated of the disease shape respectively. The combined optimal sets of textural colour and shape features is implemented in the cloud-side of



(a) (b)

Fig 6. Top-10 selected features (5-fold) from (a) 260 colour-based features from three feature selection methods (feature index (1) homogeneity, (2) contrast, (3) energy, (4) correlation, (5) sum of squares, (6) sum average, (7) sum variance, (8) sum entropy, (9) entropy, (10) diff variance, (11) diff entropy, (12-13) information of correlation measures #1&#2, (14) mean, (15) standard deviation, (16) skewness (17) kurtosis (18) coarseness (19) contrast (visual-based), (20) directionality for greyscale, (21-40) R, (41-60) G, (61-80) B, (81-100) Y, (101-120) Cb, (121-140) Cr, (141-160) H, (161-180) S, (181-200) V, (201-220) L, (221-240) a, (241-260) b colour components; and **(b) 136 shape features** (feature index (1) principal axis ratio, (2) area ratio, (3) compactness, (4) circularity, (5) orientation, (6) complexity, (7-13) 1st Hu's invariant moments, (14) equivalent diameter, (15) hydraulic radius represented by mean of shape values with AC = 0 (all disease patches), (16-30) peak index of shape histogram with AC = 0, (31-60) AC = 0.01, (61-90) AC = 0.05, (91-120) AC = 0.1, (121-122) PAR+AR, PAR+AR+SLD represented by mean value at AC = 0, (123-124) peak index at AC = 0, (125-128) AC = 0.01, (129-132) AC = 0.05, (133-136) AC = 0.1)

the system cooperating with the automatic leaf segmentation performed on a mobile phone and captured using a template overlay; initial testing shows promising results accessible from a smartphone. The top potential sets of features can be further implemented as a lightweight feature set on a standalone-mobile application.

ACKNOWLEDGMENT: I would like to thank David Gibson, University of Bristol who provided the FERA labelled images and the (EPSRC funded) India-UK Advanced Technologies Centre for partially funding the project.

6. REFERENCES

[1] M. B. Riley, M. R. Williamson, and O. Maloy, “Plant Disease Diagnosis,” in *The Plant Health Instructor*, 2002.

[2] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu, and B. Yang, “Application of Support Vector Machine for Detecting Rice Diseases Using Shape and Color Texture Features,” *2009 International Conference on Engineering Computation*, pp. 79–83, 2009.

[3] H. Wang, G. Li, Z. Ma, and X. Li, “Image recognition of plant diseases based on backpropagation networks,” in *Fifth International Congress on Image and Signal Processing*, 2012, pp. 894–900.

[4] Y. Tian, C. Zhao, S. Lu, and X. Guo, “SVM-based Multiple Classifier System for Recognition of Wheat Leaf Diseases,” in *Conference on Dependable Computing*, 2010, pp. 2–6.

[5] Z. Sarayloo and D. Asemani, “Designing a classifier for automatic detection of fungal diseases in wheat plant,” in *23rd Iranian Conference on Electrical Engineering*, 2015, pp. 1193–1197.

[6] H. C. Peng, F. H. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[7] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, “Overcoming the myopia of inductive learning algorithms with RELIEFF,” *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[8] A. N. Mucciardi and E. E. Gose, “A Comparison of Seven Techniques for Choosing Subsets of Pattern Recognition Properties,” *IEEE Transactions on Computers*, vol. C–20, no. 9, pp. 1023–1031, 1971.

[9] “The Food & Environment Research Agency.” [Online]. Available: http://fera.co.uk/.

[10] R. M. Haralick, K. Shanmugam, and DinsteinIts’shak, “Textural Features for Image Classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, no. 6, pp. 613–621, 1973.

[11] H. Tamura, S. Mori, and T. Yamawaki, “Textural Features Corresponding to Visual Perception,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 8, no. 6, pp. 460–473, 1978.

[12] P. Siricharoen, B. Scotney, P. Morrow, and G. Parr, “Automated Wheat Disease Classification Under Controlled and Uncontrolled Image Acquisition,” in *Image Analysis and Recognition*, vol. 9164, 2015, pp. 456–464.

[13] M. K. Hu, “Visual Pattern Recognition by Moment Invariants,” *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[14] D. Gibson, T. Burghardt, N. Campbell, and N. Canagarajah, “Towards Automating Visual In-field Monitoring of Crop Health,” in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 3906–3910.