International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-4, Issue-3, March 2016

Paddy Crop Disease Finder Using Probabilistic Approach

P.Tamije Selvy, Kowsalya N, Priyanka Krishnan

***Abstract*— Automatic detection of plant diseases is more important in monitoring large fields of crops, through the symptoms that appear on the leaves. This paper deals with the detection of Paddy Leaf diseases from the images captured by the farmers over their fields using smart phones via mobile application. The paddy leaf images are captured and possible attribute values like age, colour, variety, appearance and temperature are collected from the farmer. Then the noise over the image is removed, which is followed by pre-processing and segmentation using Statistical Region Merging (SRM) method. Then key features like mean, standard deviation, variance, correlation coefficient, entropy, homogeneity are extracted. The classification is performed for each image pixel using feature descriptors defined on its neighbourhood. Classification module consists of a probabilistic framework that generates the posterior probability which maps disease area estimates in the imagery data. To perform structured Pattern Prediction, Conditional Random Field (CRF) method is used. The energy minimization in CRF is implemented through α-expansion and αβ-swap algorithms. Then the training database is updated with labelled examples using expert knowledge. Then the predicted information are found from the training database with maximum optimality and informed to farmers via the mobile application**

***Index Terms*—Paddy crop diseased Leaf Images, Feature Ranking, Conditional Random Field (CRF), Statistical Region Merging (SRM), Probabilistic Neural Network (NN), Supervised Learning**

I. INTRODUCTION

Data mining, is an important part of knowledge discovery, is defined as the automated discovery of previously unknown, nontrivial and potentially useful information from databases . Database mining is the process of generating high-level patterns that have acceptable certainty and are also interesting from a database of facts. Image mining draws basic principles from concepts in databases, machine learning, statistics, pattern recognition and soft computing.

Dr. P.Tamijeselvy, CSE, Sri Krishna College Of Technology, Coimbatore, India,9843598205.

Kowsalya N., CSE, Sri Krishna College Of Technology, Coimbatore, India , 8056420073.

Priyanka Krishnan, CSE, Sri Krishna College Of Technology, Coimbatore, India, 9487118268.

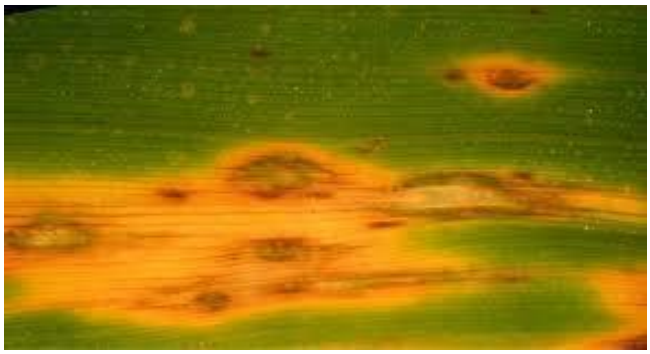

Fig 1 Diseased image

A disease that affects the plants diminishes the crop growth and productivity for a country. The diseases can be easily recognized from the visible part of the plant like leaves, stem, fruits, etc. as shown in Fig 1. Many farmers are unaware of the diseases in their own lands.

Image processing technique is one wise approach to find the disease through the images of the disease affected plant. It is further improved by several data mining algorithms by parsing through trained datasets that enumerates the disease more precisely. The Knowledge Discovery of these data helps in predicting the diseases through pattern extraction methods.

II. RELATED WORKS

The application of Data Mining techniques involves number of challenges relating to representation of images. The images can be represented variedly depending on the algorithm employed over the data.The following analyses of image processing techniques and data mining algorithms helps in improving the design of the proposed system.

Arngren .M et al., (2011) proposed the analysis of pre germinated barley using hyper spectral image analysis [19]. Agricultural productivity is strongly dependent on monitoring and planning activities which results in the production estimation and land use to enhance agricultural activities. The algorithm used in the system is:1)Remote Sensing Image Classification,2)Genetic Programming,3)Optimum-Path Forest,4)Relevance Feedback. A new hybrid method, named GOPF, which uses a GP framework to create composite image descriptors and the optimum-path forest (OPF) classifier to determine regions of interest. The classification method represents each class of objects by one or multiple optimum-path trees rooted at key samples called prototypes. A fitness values for each individual is used to assign the based on the ranking of the training set and this value is used to select the best individuals. The system uses image descriptors to encode user‟s relevance feedback. GOPF has presented good results with respect to the identification of pasture and coffee crops.

Yu, q., Gong et al., (2006) proposed the framework for Object-Based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery [18]. Various spatial resolution images were used to aggregate data in order to reduce the variation within the object, and minimize the classification error. The variability within an object can provide additional information that can be used for image classification and group of pixels that form image segments are called patches. Several methods are used in this spectral imagesuchas,1)ImageSegments,2)Classification,3)Object-oriented approach,4)Multivariate analysis.

In the case of the object-based classification, patches are not expected to consist of pixels with completely homogeneous spectral radiances, but rather certain levels of variability are expected. Within a patch, pixels from the outliers of the class distribution are likely to be misclassified. The window based approaches use arbitrary groupings and return the value of the window to the central pixel in the image set. The variation in an object is used as one of the characteristic of the object in this method, whereas it is an obstacle with traditional pixel-based classification methods. The object-based approach resulted in a pleasing of spatial structure compared to the noisy patterns of traditional pixel based classification. The highest accuracy 0.783 was obtained from the classification method. Anibal Gusso et al., proposed the Algorithm for Soybean Classification using Medium Resolution Satellite Images [17].An algorithm for soybean classification was developed as an objective automated tool for the reliable calculation of real crop parameters as to yield, production and other important to decision-making polices. A set of 39 municipalities was analysed for eight crop years between 1996/1997 and 2009/2010. RCDA estimates were compared to the official estimates of the Brazilian Institute of Geography and Statistics (IBGE) for soybean area at a municipal level. The soybean classification procedure was Reflectance-based Crop Detection Algorithm (RCDA) which uses lower limits of the spectral range (lower reflectance values) for each band in the image set. The system uses various methods such as:1)Physically Driven Components (PDC),2)MODIS Crop Detection Algorithm,3)Leaf Area Index,4)Global Land Survey. The overall map accuracy was 91.1% and its Kappa Index of Agreement was 0.76.

The diseased images can be easily analysed, processed through various image processing techniques. Since the diseased part is isolated from the leaf images, colour based segmentation chosen. For feature extraction neighborhood pixels should be manipulated for obtaining the features.

the features are extracted, the pixels values are plotted in a matrix form which enables the easy identification on diseases. Images should be classified based on the feed forward approach, where unsupervised learning is followed. To attain these goals, techniques like

1.Statistical Region Merging for Image Segmentation

2.Co-Occurrence Distribution Matrix Calculation for Feature Extraction

3.Probabilistic Neural Network Approach for Classification

4.Conditional Random Field for Pattern Recognition

are chosen. These approaches helps in identifying the disease without any deviation thus overcoming the drawbacks which were analysed on the related works of few systems discussed above.

III. Proposed Scheme

The overall scheme is depicted in figure 2 which includes image pre-processing, image segmentation, feature extraction, classification and pattern recognition.

3.1 Image Pre-Processing

The pre-processing of images helps in removing the noise and brings uniformity over the images. It is the processing of image using mathematical operations by using any form of signal processing for which the input is an image, a series of images or a video; the output of image processing may be either an image or set of characteristic or parameters related to the image.

Removing abnormality can widely improve the linearity towards finding the disease. Since the images are collected from various fields of various resolutions image rectification and image restoration is done.

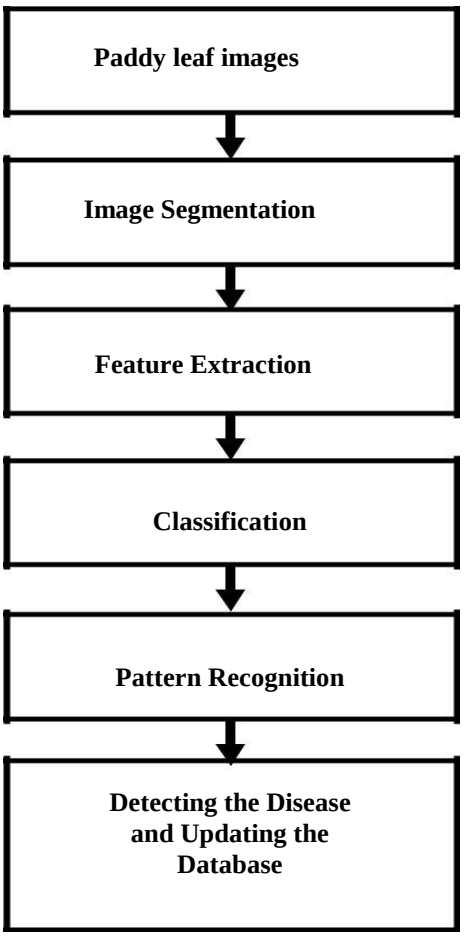

Fig. 2 Overview of the system

3.2 Image Segmentation

Image segmentation is crucial step for the object-based remote sensing imagery analysis. . A segment can be considered to be any region having pixels with uniform spectral characteristics [3]. The aim of segmentation is to find regions with uniform values for the different spectral bands

representing a particular land-cover class. The closely occurring regions are segmented such that the disease can be differentiated from the healthy part of the leaf which is simulated in the Fig 3.

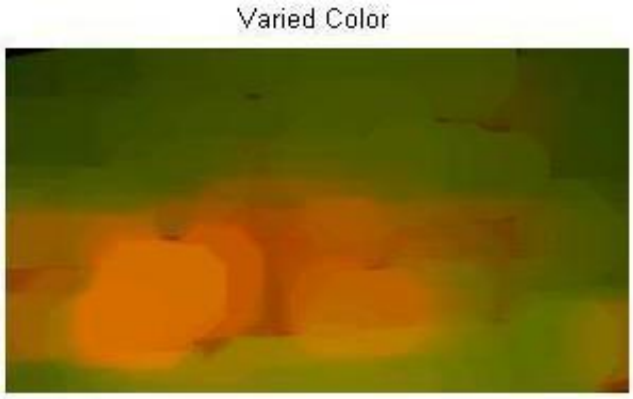

Fig 3Varied Colour Image of the Diseased Leaf Image

3.2.1 Statistical Region Merging

The SRM algorithm initially considers each pixel as a region and merges them to form larger regions based on a merging criterion [5]. The merging criterion is the red, green, blue, and NIR values of neighbouring pixels that correspond to dR, dG, dB, and dNIR, respectively, merges two regions if (dR < threshold & dG < threshold & dB < threshold & dNIR < threshold). The merging criterion can be formalized as a merging predicate that is evaluated as true if two regions are merged and false otherwise.

Input: Pre-processed image **Output:** Segmented PlotRegions

For the chosen the image, n number for segmentations are generated. Image Gradient is computed by filtering X and Y value co-ordinates, which is manipulated using Formula 1

Imggrad = sqrt(+) **(1)**

Build two regions of images and map those regions. Two figures are obtained, figure 1 is the segmentation map and figure 2 is the segmentation map with the average colour in each segment. Generate the Cluster list for X and Y co-ordinate values. Colour the image based on the segmentation and display the varied colour image.

3.3 Feature Extraction

The availability of large number of spectral features in diseased data should make it possible to discriminate more disease prone area with greater accuracy [4]. Unfortunately, many available bands of spectral data are highly correlated and provide redundant information as most of the surfaces shared similar spectral characteristics in continuous bands. In order to realize the potential of the high dimensional data, it is desirable to select the optimum subset of channels using the plotted matrix value for analysis and parameter, as well as to reduce computational requirements. Feature extraction is performed using Co-Occurrence Distribution Matrix (CDM) from the required plot in the given dataset using Feature Ranking Algorithm.

3.3.1 Co-Occurrence Distribution Matrix

The Co-Occurrence Distribution Matrix is used to find the nearby pixel range. An 8×8 matrix is used to plot the image values over the matrix. Mean, Standard deviation, Variance, Entropy, Correlation Coefficient are calculated through the matrix values [11].

Input: Segmented regions of crop fields.

Output: Features include parameters such as mean, variance, Standard deviation, Covariance for several criteria.

Co-occurrence matrix **C** is defined over an **n × m** image **I.** Parameterized by an offset **(Δx,Δy)**, manipulated using Formula 2

$$C_{\Delta x,\Delta y}(i,j)=\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \text{ and } I(p+\Delta x,q+\Delta y)=j\\ 0, & \text{otherwise}\end{cases} \quad \textbf{(2)}$$

where *i* and *j* are the image intensity values of the image, *p* and *q* are the spatial positions in the image **I** and the offset **(Δx, *y*)** depends on the direction used $\theta$and the distance at which the matrix is computed *d*. Plotted in 8×8 matrix where values are extracted, with the extraction terms.

Mean - an average of n number of pixel values from the matrix, where the mean value for image is calculated.

Variance - measures the farthest and nearest pixel value in the matrix from the mean.

Standard Deviation - the amount of variation or dispersion of a pixel values in the matrix.

Covariance - the mean value, calculated using each pair of pixels.

Entropy - Statistical measure of randomness that can be used to characterize the texture of the image.

Homogeneity - Returns a value that measures the closeness of the distribution of -pixel value.

Correlation Coefficient – Sample correlation coefficients for a pixel matrix are generated.

3.4 Classification

The intent of the classification process is to categorize all pixels in a digital image into one of several classes [16]. The objective of image classification is to identify and portray, as a unique grey level, the features occurring in an image in terms of the object or type of leaf. Two main classification methods are Supervised Classification and Unsupervised Classification. A feed-forward approach is used to classify the datas based on the probabilistic approach. It consists of a Probabilistic Neural Network (NN) framework that generates the posterior probability maps of the disease area estimates from images.

3.4.1 Probabilistic Neural Network

To understand the basis of the Probabilistic Neural Network paradigm, it is useful to begin with a discussion of the Bayes decision strategy and nonparametric estimators of probability density functions. It will then be shown how this statistical technique maps into a feed-forward neural network structure typified by parallel simple processors [12]. **Input:** Features from the image.

Output: Classified datas.

Classification is performed for each image pixel based on the feature descriptors. A neighbourhood system for a pixel p is a set Πp defined as in Equation 3

Πp = U(rL−pL≤τ) r. **(3)**

Here, rL and pL are the locations, i.e. the ordered tuple(x,y) for pixels r and p, respectively, where x is the X-coordinate (along the horizontal laxis),and y is the Y-coordinate(along

the vertical axis), i.e., rL−pL=δ if rL lies on a δ×δ window centred at pL.

The activation function is tan sigmoid(tan hyperbolic) forbidden layers and linear for output layer as in Formula 4

$$\sigma(t) = \tan h(t) = \quad -/ \quad + \qquad \textbf{(4)}$$

The outputs of an NN trained by minimizing the MSE function approximate the conditional averages of the target data as in Formula 5

$$\Psi ij(xi,x\,j) = \{ \qquad \textbf{(5)}$$

where tk"s are the set of target values that represent the class membership of the input vector xk, and p(tk|x) is the probability that the input vector x attains the target value tk. The neighbourhood system for pixel pL is shown in the Fig. 4

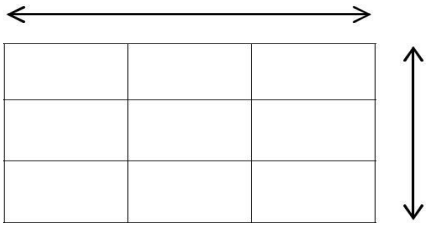

Fig. 4 Neighbourhood system for pixel pL, where rL −pL = δ.

3.5 Pattern Recognition

Pattern recognition focuses on extracting patterns and regularities in data. Pattern recognition has higher interest to formalize, explain and visualize the pattern [13]. They are in many cases trained from the supervised learning with labels, but when no labeled data are available, unsupervised learning is used to train the data sets.

3.5.1 Conditional Random Field

A conditional random field (CRF) has been used in the pattern recognition literature for performing structured prediction. In structured prediction, the labelling of a pixel depends not only on the feature values of that particular pixel but also on the values assumed by neighbouring pixels.

Input: Classified datas

Output: Segmented plotted region

A CRF defines a set of random variables XC conditionally dependent on each other as a clique c. A probability distribution associated with any random variable Xi of a clique is conditionally dependent on the distributions of all other random variables in the clique. The objective function takes the form as in Equation 6

$$E(x) = \sum \qquad + \sum \qquad + \sum \qquad \textbf{(6)}$$

where ψi (xi) is the unary potential, ψij(xi,x j) is the pairwise potential, and ψc (xc) is the function associated with higher order region consistency potential defined over segment S.

The unary potential term is defined as Formula 7

$$\Psi i\ (xi) = \theta N \psi N(xi) + \theta band \psi band(xi) \quad \textbf{(7)}$$

which denotes the potential due to the output produced by the NN classifier.

The potential derived from classifier output as in Formula 8

$$\psi N\ (xi) = -\log P(ci/x) = -\log yi \qquad \textbf{(8)}$$

Similarly, the pairwise term ψij(xi, xj) is updated to encode the band information in Mathematical Equation 9

$$\Psi ij(xi,x\,j) = \{$$

$$| \qquad | \textbf{(9)}$$

3.5.2 CRF Algorithm

Two CRF learning algorithm is used to get arbitrary value for the image.

Algorithm1
α-Expansion Algorithm

1: procedure ALPHAEXPANSION
2: Assign an arbitrary labelling y to the pixels of the image.
3: done← 0
4: for each label α ∈ L do
5: find y←argminE(y) among y where y lies within one α-expansion of y
6: if E(ˆ y) <E (y) then
7: done← 1
8: if done =1then
9: goto3.
10: return y.

Algorithm2
αβ-Swap Algorithm

1: procedure ALPHABETASWAP
2: Assign an arbitrary labelling y to the pixels of the image.
3: done← 0
4: for each pair of labels α,β ∈ L do
5: find ˆ y ← argminE(y) among y where y lies within one αβ-swap of y
6: if E(ˆ y) <E (y) then
7: y ← ˆ y
8: done← 1
9: if done =1then
10: goto3.
11: return y.

3.6 Disease Detection and Updating the Database

Once the results are obtained training database is updated with the user obtained attribute values, segmented images, feature value, accurate classifiers and random potentials. The data were accessed through a MySQL database.

The database is called „Trained set". The trained set contains all the necessary to features to find the disease.

Disease database for the paddy crop is devised where the disease properties like disease name, size, shape, colour, feature values are stored. Some of the common diseases like Rice blast, Blight or Brown spots, Bunt of Rice, Bacterial Blight, Bakane Disease, Ufra of Rice, Stem Rot are studied and their symptoms, perpetuation and control are stored. It has been observed that the output accuracy varies with respect to paddy diseases.

To measure the performance of a paddy disease test, the concepts sensitivity and specificity are often used. Say some of the test leaves have disease and it is called true recognition (TR) if the system recognizes the disease properly.

In addition, if the system provides misleading results then it is called false recognition (FR). This accuracy chart is stored alongside with the diseases. By comparing the accuracy classifier the diseases can be found and predicted to farmers. The diseases are informed through mobile application.

IV. EXPERIMENTAL RESULTS

4.1 Segmentation

In the segmentation method, Table 1 depicts the comparison of Existing Segmentation methods.

Table 1 Comparison of Segmentation Techniques

Algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)
Seeded Region Growing	74.5	77.6	67
USCT Region	77.3	81	72
HS	83.2	80.3	75
Marker based HSEG	85	79	70
SRM	88	85.3	78

Fig. 5 shows the pictorial representation of the Comparison of Segmentation techniques.

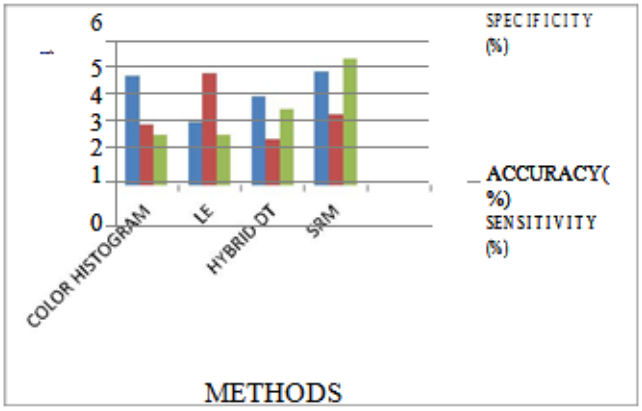

Fig. 5 Comparison of Segmentation Techniques

4.2 Feature Extraction

In the Feature Extraction method, Table 2 depicts the Old Feature Extraction methods comparing with the proposed Co-Occurrence Distribution Matrix

Table 2 Comparison of Feature Extraction Techniques

Algorithm	Accuracy (%)	Specificity (%)	Sensitivity (%)
GNN	79.1	75	71.1
SURF	83.5	73.1	69.9
SHIFT	81.2	77	74.3
CDM	87.3	84.2	80.2

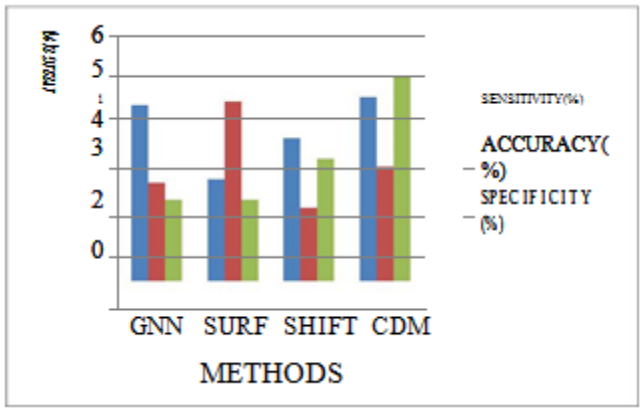

Fig. 6 Comparison of Feature Extraction Techniques

V. CONCLUSION

The image processing system enhances the input image and segments the images based on the pixel and texture over the leaves. Using these texture value features are extracted from which feature ranking of several feature values are obtained. The feature values helps in finding the disease of the paddy in further modules. This system can be enhanced to be used for the disease detection of all kind of crops. The images are captured by the farmers through smart phones hence the mobile application is also developed.

REFERENCES

[1] X. Sun, H. Wang, and K. Fu, "Automatic detection of geospatial objects using taxonomic semantics," IEEE Geosci. Remote Sens. Lett., vol. 7, no. 1, pp. 23–27, Jan. 2010.

[2] A.C.Holt, E.Y.W.Seto, T.Rivard, and P.Gong, "Object-based detection and classification of vehicles from high-resolution aerial photography,"Photogramm. Eng. Remote Sens., vol. 75, no. 7, pp. 871–880, 2009.

[3] E. Sharon, A. Brandt, and R. Basri,"Segmentation and boundary detection using multiscale intensity measurements," in Proc. IEEE CVPR, 2001, pp. 469–476. [Online]. Available: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2001-1.html#SharonBB01 5706 IEEE

TRANSACTIONSON GEOSCIENCE AND REMOTESENSING, VOL. 53, NO. 10, OCTOBER 2015

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern

Anal. Mach. Intell., vol. 22, no. 8, pp. 888– 905, Aug. 2000.

[5] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu,

"Image parsing: Unifying segmentation, detection, and recognition," in Proc. Int. J.

Comput. Vis., 2003, vol. 1, pp. 18–25.

[6] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines,"

IEEE Geosci. Remote Sens. Lett., vol. 4, no. 4, pp. 674–677, Oct. 2007.

[7] R. Komura, M. Kubo, and K.-I. Muramoto,

"Delineation of tree crown in high resolution satellite image using circle expression and watershed algorithm," in Proc. IEEE IGARSS, Sep. 2004, vol. 3, pp. 1577–15.

[8] L. Duncanson, B. Cook, G. Hurtt, and R. Dubayah, "An efficient, multi-layered crown delineation algorithm for mapping individual tree structure across multiple ecosystems," Remote Sens. Environ., vol. 154, pp. 378–386, Nov. 2014.

[9] B. D. Cook et al., "NASA goddards LiDAR, hyperspectral and thermal (G-LiHT)
airborne imager," Remote Sens., vol. 5, no.8, pp. 4045–4066, 2013. [Online].Available: http://www.mdpi.com/2072-4292/5/8/4045

[10] T.-W.Chen, Y.-L.Chen, andS.-

Y.Chien, "Fastimage segmentation based onk-means clustering withhistograms inHSVcolor space"inProc.IEEE Signal

Process. Soc. MMSP, 2008, pp. 322–325.

[11] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEETrans. Syst., Man

Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973. [33] L. kiat Soh and C.

Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," IEEE Trans. Geosci. Remote

Sens., vol. 37, no. 2, pp. 780–795, Mar. 1999.

Available: http://www.mdpi.com/2072-4292/5/8/4045

[12] T.-W.Chen, Y.-L.Chen, andS.-

Y.Chien, "Fastimage segmentation based onk-means clustering withhistograms inHSVcolor space"inProc.IEEE Signal

Process. Soc. MMSP, 2008, pp. 322–325.

[13] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEETrans. Syst., Man

Cybern., vol. SMC-3, no. 6, pp. 610–621, Nov. 1973. [33] L. kiat Soh and C.

Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," IEEE Trans. Geosci. Remote

Sens., vol. 37, no. 2, pp. 780–795, Mar. 1999.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.

Salakhutdinov, "Improving neural networks by preventing coadaptation of feature detectors," unpublished paper. [Online].
Available: http://arxiv.org/abs/1207.0580

[15] C. M. Bishop, Neural Networks for Pattern Recognition. New York, NY, USA: Oxford Univ. Press, 1995.

[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. 8th

ICML, San Francisco, CA, USA, 2001, pp. 282–289.

[17] G. H. Bakir et al., Predicting Structured Data (Neural Information Processing). Cambridge, MA, USA: MIT Press, 2007.

[18] Gusso, A., & Ducati, J.R.(2012). Algorithm for Soybean Classification Using Medium Resolution Satellite Images. Remote Sensing, 4(10), 3127-3142.

[19] Yu, Q., Gong P., Clinton N., Biging G.,Kelly M., & Schirokauer D.(2006).Object- Based Detalied.

.