# Tomato Plant Disease Classification for Mobile Phone Image Using SIFT-Beta Feature and Color Statistical Feature

Chit Su Hlaing
University of Computer Studies, Mandalay (UCSM)
Mandalay, Myanmar
chitsuhlaing@ucsm.edu.mm

Sai Maung Maung zaw
University of Computer Studies, Mandalay (UCSM)
Mandalay, Myanmar
saimaungmaungzaw@ucsm.edu.mm

## Abstract

*Plant disease classification is essential for food productivity and disease diagnosis in agricultural domain. The probability distribution and statistical properties are essential in image processing to define the features of typical image. The general usage of (Scale Invariant Feature Transform) SIFT has local feature extraction and global feature extraction (bag-Of-Features approach) for classification, and its classification result for unknown data also depends on code book (global feature) generation. Instead of using bag-Of-Feature approach, we proposed to apply Beta probability distribution model for SIFT to be directly represent the image information and then formed SIFT-Beta. The color statistics feature is extracted from RGB color space and then combines with SIFT-Beta to produce proposed features. The proposed feature is applied in Support Vector Machine classifier. The classifier is trained for seven labels of tomato with six diseases and healthy.*

## 1. Introduction

The disease of Fruits and vegetable crop disease is an important issue for the damage of fruit and vegetable crops. Fruits and vegetables are classified from both a botanical and culinary standpoint. Fruit and vegetable growing are the important and paying branches of horticulture. Well maintained and growing offer many economic advantages for agriculture development. In agriculture domain, computer vision and digital image processing are successfully applied to develop the farmer life. A good probability distribution model can be applied to represent the image information effectively in digital image processing.

There are many development and advance research for plant disease classification based on digital image processing.

The digital image processing with deep learning and population of smart phone are combined to support food production. A deep convolutional neural network was proposed for PlantVillage dataset to classify 26 diseases with 14 crop species. Their classification accuracy reached 99.35% withheld-out test set. In PlantVillage dataset, all of the images are leaves image with diseases and healthy for 14 crop species. All of the images are taken by smartphones with different resolution and different image sizes. PlantVillage images are available in their public web site to support researchers in image processing and machine learning. [2].

H and I3a and I3b color space are introduced for preprocessing step of plant disease classification. The color conversation from RGB to proposed color space is performed at the preprocessing step for image enhancement. And feature extraction is performed for banana, maize, alfalfa, plantain, cotton and soya leaf disease classification. The banana leave images in the dataset are obtained from the Universities and research institutes from USA. Image labeling is provided by International Network for the Improvement of Banana and Plantain crops. [1].

A set of statistical features was introduced to overcome the challenges of agriculture. They applied proposed feature for beet leaves cellphone images with six diseases called Uromyces betae, Ramularia beticola, Cercospora beticola, the bacterium Pseudomonas and Phoma betae caused by fungi. They used two datasets called FULL and STUDY that are consist of cell phone camera images of a beet leaf suffering from above these six diseases. STUDY dataset consists of 495 leave images with six labels of beet diseases. FULL dataset consists of 2957 leave images for six types of beet diseases. All of images from both datasets are taken by the cellphone camera with different resolution and disease region exits at the center of the images and no background and noise data. The images only contain whole leaf region full with camera. The labeling of all images are decisions of the German Federal Office for Agriculture and Food. [5].

Radial Basis Function Neural Network (BRBFNN) was proposed to identify a plant disease

on a leaf. Bacterial aging optimization is used to assign optimal weight of the network for plant disease on leaf classification. The dataset consist of plant disease on leave images for six different types of fungal diseases: late blight, leaf spot, early blight, common rust, leaf curl, and cedar apple rust. For a validation, they had two set of dataset, one for segmentation experiment and other for plant disease classification. The first one has 6 images from planet natural with six types of fungal diseases and the second has 270 images from PlantVillage dataset for six types of diseases. [7].

The different models of convolutional neural network were proposed to use for plant disease classification and identification. An open database of over 80000 images, with 25 crops for 58 class labels for (plant, disease) was used for experimental performance. All of the images in database are mobile camera images of plant leaf belonging with healthy and disease regions. And images have Field conditions and Laboratory conditions for training and classification. [8].

The different types of image processing techniques were proposed to solve the problems in defining fungal disease symptoms affection on plant leave for fruit and vegetable of plants. A framework for different type of fruit and vegetable is proposed to develop the quality and quantity of food production by reducing diseases on plant. They proposed to use different type of features and classifiers according to the type of crops. The dataset consists of fungal diseases images of fruit, vegetables, commercial and cereal crops. All of the images are collected from University of Horticultural Sciences (UHS) and University of Agricultural Sciences (UAS). [6].

The properties of SIFT features rotation invariant, scaling invariant and illumination invariant. In calculation, the matrix structure of SIFT feature is too difficult to handle and has multiple steps for global feature representation. This paper focus on probability distribution to be directly address the SIFT feature of an image. The main contribution is to model SIFT feature using Beta distributions to form SIFT-Beta is. Hence, the proposed feature consists of color statistics feature and SIFT-Beta feature for classifying seven labels in tomato disease classification. The proposed feature supports to classify the seven labels of tomato (six diseases and healthy) from PlantVillage. The organization of the paper has 4 main sections: proposed methodology is shown in in Section Two. In Section 3 and Section 4, the experimental results and conclusion are presented.

## 2. Proposed Methodology

Our plant disease classification system has three main parts

1. Image preprocessing,
2. Proposed feature extraction
3. Training and disease name labeling.

The proposed feature is applied in SVM classifier to classify seven labels with six diseases and healthy of tomato leaves shown in Figure 1(a) to (g).

### 2.1. Image Preprocessing

The input image is tomato leaf disease of RGB color image in jpeg format. The preprocessing stage has four steps: green region detection, disease region detection, whole leaf region detection and region filling.

a) **Green region detection**. To extract green region from input image, separate the RGB color channel of the image into red channel, green channel and blue channel. By using all three channels of images and conditions, red and blue value must less than green value. The output of this stage is presented in figure 2 (b).

b) **Disease region detection.** To extract disease region from input image, separate the RGB color channel of the image into red channel, green channel and blue channel. By using all three channels of images and conditions, the blue and green channel value must be less than the red value. The output of this stage is presented in figure 2 (c).

c) **Whole leaf region detection**. To extract whole leaf area, bit OR operation is performed on the green region and disease region of whole leaf image. The output of this stage is presented in figure 2 (d).

d) **Region filling.** The median filtering and region filling are used to remove salt and pepper noise and to fill holes in an image. An output is presented in figure 2 (e).

i ii iii iv v vi vii

Figure 1. **Tomato Leave Images (a) Leaf Mold, (b) Two Spotted Spider Mite, (c) Septoria Leaf Spot, (d) Bacterial Spot, (e) Late Blight (f) Target Spot and (g) healthy**

## 2.2. Proposed Features Extraction

The preprocessing is performed on input leave images to remove unwanted background information and then extract SIFT feature on preprocessed images. Beta probability distribution is used to model the extracted SIFT feature to form SIFT-Beta. The color statistics is also extracted from preprocessed image. And then texture feature called SIFT-Beta is combined with color feature called color statistic to form proposed feature. The proposed feature considers the significant of texture and color to classify tomato disease name. The scheme of proposed feature extraction is presented in Figure 3.

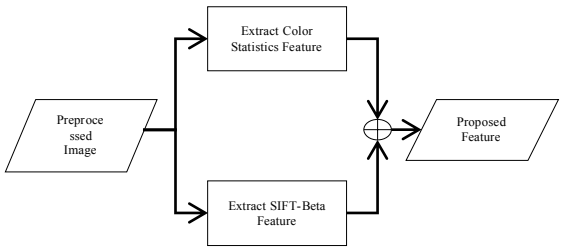
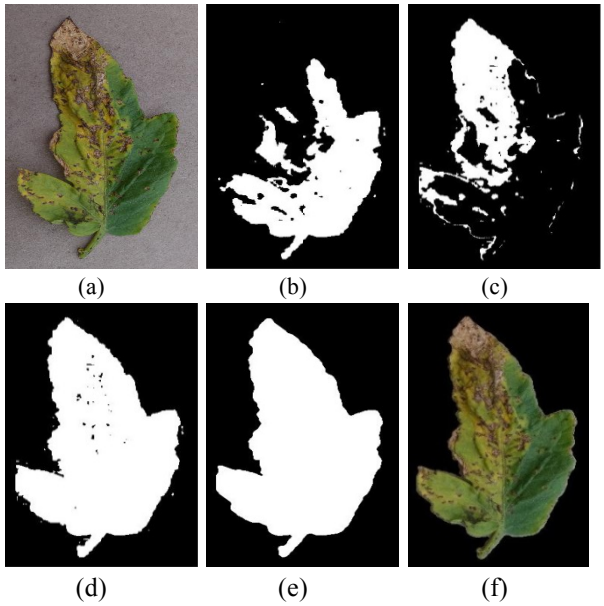



**Figure 3.Scheme of proposed feature extraction.**

### 2.2.1. SIFT Texture Feature

The Scale Invariant Feature Transform (SIFT), feature is specially applied in image retrieval and image matching. This algorithm has main two parts: keypoint detection and feature descriptor building. The keypoint detection produces location information of keypoint by using scale space Difference-Of-Gaussian (DOG). And descriptor building produces the feature vector value for keypoints using keypoints location and Gaussian window. The SIFT algorithm produces descriptor and location data for interest points of an image.

### 2.2.2. Beta Probability Distribution Model



(a) (b) (c)

(d) (e) (f)

Figure 2. **Visual representation for image preprocessing (a) Input Tomato Leaf, (b) healthy region extracted image, (c) diseased region extracted image, (d) Performed bit OR operation on the green region and disease region , (e) Region filled image, (f) Background removed image**

In Beta Probability Distribution model, maximum likelihood estimation algorithm is used to estimate parameters of the Beta probability distribution of SIFT feature. The Beta distribution has the relation with normal distribution. The Beta probability distribution function is:

$$f(x) = \frac{(x-k)^{(a-1)}(\sigma-x)^{(b-1)}}{B(a,b)(\sigma-k)^{a+b-1}} \quad (1)$$

where k and σ are the lower and upper bounds, respectively, $k \le x \le \sigma$ , a and b are the shape parameters greater than 0, and B(a, b) is the beta function. The general formula of Beta is:

$$B(x,y) = \int_0^1 s^{x-1}(1-s)^{y-1}\,ds \qquad (2)$$

Using the distribution of the maximum order statistics, maximum likelihood estimation (MLE) algorithm produces the parameters of the Beta distribution of SIFT features. If the set {x1, x2… xn} (SIFT feature) are independent and identically distributed from a Beta distribution, then the log-likelihood function for a sample of n observations {x1, x2… xn} is

$$\ln[L(\theta|x) = -n\ln(\sigma) + \sum_{i=1}^{n}[\left(\frac{1}{k}-1\right)\ln(y_i) - (yi)^{1/k}] \quad (3)$$

where θ = (k, σ, μ) and yi= [1-(k/ σ) (x)]. The MLE of k and σ can be identified by solving the equation (4).

$$\frac{1}{\sigma}\sum_{i=1}^{n}\left[\frac{1-k-(y_i)^{1/k}}{y_i}\right] = 0 \qquad (4)$$

The flow of SIFT-Beta feature extraction on RGB and gray scale color spaces is presented in Figure 4.

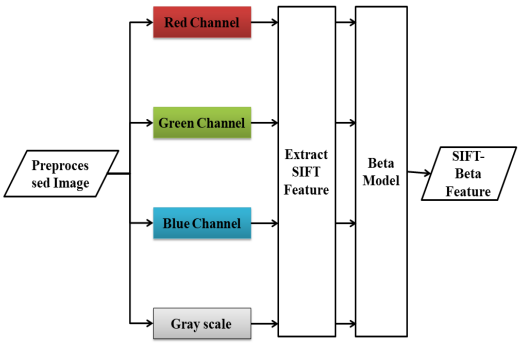



**Figure 4. Scheme of SIFT-Bata feature extraction.**

### 2.2.3. Color Statistics Feature

The color statistics is extracted from preprocessed image that represents in numerical vector form. Color statistics feature is calculated from mean, standard deviation and moments degrees. The color statistics feature is calculated on RGB color space of the preprocessed image. The moment of degree 'n' is:

$$M_h(n) = \frac{\sum_{j=1}^{N}(ch_j - mean_{ch})^n p(ch)}{\sigma^n} \qquad (5)$$

where $mean_{ch}$, σ are mean and standard deviation of ch, and p(ch) the marginal histogram of ch , which h been estimated through a 256-bin histogram, N is the number of elements in spectral dimension h and ch is the value of element in index j which is calculated according to the structure of bins. It is also calculated the moment order up to five since higher orders are not effectively changed the classification accuracy. [9].

### 2.2.4. Support Vector Machine Classifier

Different types SVM classifiers according to kernel functions are linear SVM classifier, quadratic SVM classifier, cubic SVM classifier and fine Gaussian SVM classifier. The quadratic SVM classifier is used to classify seven labels of tomato diseases. Multiclass Support Vector Machine classifier was used to directly solve the multi-class problem in a single optimization process. The two main parameters of SVM classifier that can affect the construction of classifier are Box constraint level and Kernel scale. Box constraint level associated with margin-violating observations and supports to avoid over fitting. Kernel scale parameter associated with how spreads out our data point are.

### 2.2.5. 10-Fold Cross Validation

10-fold cross classification task was executed to assess the benefits of proposed feature in tomato plant disease classification. The validations are performed ten times in 10-Fold-cross validation. In every validation times, the database is divided into ten subgroups in which nine subgroups for training and one subgroup for testing. The classification accuracy is calculated in every validation times. Finally prediction speed, training time and average classification accuracy are calculated over these ten validation times.

## 3. Experimental Results

We perform state of feature comparison over 10-fold cross validation with PlantVillage Dataset and GOF test for SIFT feature and its distribution models.

### 3.1. K-Fold Cross Validation

In k-fold cross validation, the dataset is divided into k subsets, and the validation is repeated 10 times. Each validation time, one subset is used as the test set and the other k-1 subsets are put together to form a training set. Then the average classification accuracy, training time and prediction speed across all k Ts are computed. The true positive rate, false positive rate, are also calculate over this k-fold cross

validation to evaluate the classifier performance. In this research, the different values of k are used to evaluate the classifier performance. The different values of k are 2, 5 and 10.Since k-fold cross validation is bias free validation method; our evaluation for plant disease classification excludes theoretically derived values. The illustration of 10-fold cross validation for classifier performance evaluation is shown in Figure 5.


**Figure 5. 10-Fold cross validation.**

### 3.1. PlantVillage Dataset

**TABLE I.TOMATO PLANT DISEASE IMAGES FROM PLANTVILLAGE DATASET**

| Tomato Disease Name | Number |
|---|---|
| Healthy | 625 |
| Septoria Leaf Spot | 552 |
| | |
| Leaf Mold | 310 |
| Late Blight | 469 |
| Target Spot | 485 |
| Bacterial Spot | 660 |
| Two Spotted Spider Mite | 373 |
| **Total Images** | **3474** |

PlantVillage is an open database of more than 7,000 images for health and crop diseases. In this database, several cell phone cameras are used to capture the plant disease images. It is originating and developing in order to develop the machine learning algorithms and digital image processing to classify crop diseases on a smartphone [2]. Tomato plant disease images are manually downloaded from PlantVillage site. The number of images of the seven labels is shown in Table 1.

**TABLE II. CLASSIFICATION ACCURACY ACCORDING TO DIFFERENT TYPE OF CLASSIFIER**

| Classifier | Average Classification Accuracy |
|---|---|
| LDA | 74.3 |
| KNN | 62.5 |
| **SVM** | **77.1** |

In Table 2, SVM classifier has the 77.1% of average classification accuracy while LDA and KNN have 74.3% and 62.5% respectively. The SVM classifier is selected to use in this plant disease classification according to the classifier analysis results as shown in Table 2.

The important of SVM classifier is kernel function. To select kernel function for SVM classifier, we also measure average classification accuracy of SVM classifier according to different type of kernel function in the structure of 10-fold cross validation on tomato dataset with proposed feature.

**TABLE III. Classificaiton accuracy according to different kernel funciotns of svm**

| SVM (Kernel function) | Average Classification Accuracy |
|---|---|
| Linear SVM | 82.2 |
| Quadratic SVM | **86.7** |
| Cubic SVM | 84.9 |
| Fine Gaussian | 32.5 |
| Medium Gaussian | 82.4 |
| Coarse Gaussian | 83.7 |

In this plant disease classification system, the quadratic SVM classifier is used according to the average classification accuracy results in Table 3.

### 3.2. State of Art Comparison

The state of art comparison is performed with the following color texture feature vectors:

**Average color differences [11]:** one of color texture feature using average color difference to estimate the color difference between pixels. It is calculated on Lab color channel and has rotation invariance. The length of feature vector is 50.

**Normalize color space representation [10]:** one of color texture feature using preliminary dimensionality reduction of color space. The original RGB image was converted into a complex number matrix after discarding color channel by using rang/avg parameter. The texture feature is calculated by the normalized energy distribution and Gabor filter bank. It has rotation invariant property with vector length 32.

**Proposed Feature**: one of color texture feature using the combination of SIFT-Beta and color statistics with feature vector length 23.

In our experiment the comparison of state of art feature vectors are performed with following setting:

- the quadratic SVM classifier
- the box constraint level is one and
- the kernel scale is auto

In this state of art comparison, the proposed feature has highest classification accuracy over different kind of color texture features. The state of art comparison results is presented in Table 4. In Table 4, VL is the vector length, CP is the color space, CA is the average classification measured in percentages (%), TT is the training time measured in seconds (secs) and PS is the prediction speed measured in (obs/sec) over accuracy over 10-fold cross validation.

**TABLE IV. STATE OF ART COMPARISON WITH 3474 TOMATO DISEASE IMAGES OVER 10-FOLD CROSS VALIDATION AND QUADRATIC SVM CLASSIFIER**

| Feature | VL | CP | CA (%) | TT (secs) | PS (obs /sec) |
|---|---|---|---|---|---|
| Normalize color space repress[10] | 32 | P1P2 | 73.8 | 31.60 | 7000 |
| Avg color differences [11] | 50 | Lab | 77.4 | 156.9 | 6100 |
| **Proposed Feature** | 23 | Gray scale & RGB | **86.7** | 77.706 | 14000 |

### 3.3. Goodness of Fitting Tests

We measure goodness of fitting test to prove that how Beta probability distribution fit with SIFT texture feature. Kolmogorov-Smirnov GOF test was used to find out the fit probability distribution for SIFT texture feature. The value of D depends on the maximum vertical difference between the theoretical and the empirical cumulative distribution function, the Kolmogorov-Smirnov statistic (D) is:

$$D = n\left(F(x_i) - \left[\frac{i-1}{n}\right], \left(\frac{1}{n}\right) - F(x_i)\right) \geq i \geq \max \quad (6)$$

Assume that we have a random sample x1... xN from some distribution with Fn(x) cumulative distribution function. The ECDF is:

$$Fn(x) = (1/N). [\text{Observations No: } \leq x] \quad (7)$$

The GOF values of Beta, Johnson SB, Generalized Extreme Value and Generalized Pareto are calculated because their empirical histogram shapes are similar to the SIFT texture descriptor histogram shape. Kolmogorov-Smirnov test statistics (D) value is between 0 and 1. The best GOF value is near to zero and under 0.5 ($0.5 \geq GOF \geq 0$). We choose the Beta probability distribution and Generalized Pareto because these have minimum value of (D) value at rank position one and three. The calculation of the GOF values for SIFT descriptors of preprocessed image is shown in Table 5.

**TABLE V. KOLMOGOROV-SMIRNOV STATISTIC (D) FOR SIFT DESCRIPTORS OF TOMATO IMAGE AND 4 DIFFERENT PROBABILITY DISTRIBUTIONS**

| No. | Distribution | D |
|---|---|---|
| 1. | Beta | **0.149** |
| 2. | Johnson SB | 0.159 |
| 3. | Generalized Extreme Value | 0.172 |
| 4. | Generalize Pareto | 0.181 |

## 4. Conclusion

The tomato plant leaf image captured by mobile phone's camera is the input and plant disease is the output in our tomato plant disease classification system. We proposed a set of statistical feature based on color and texture feature. In our proposed methodology, we consider whole leaf region to get the exact image information for feature extraction. The proposed reached 86.7% of classification accuracy features with quadratic SVM classifier. We also showed that SIFT feature fit with Beta distribution by calculating Goodness of fitting (GOF) result. Our proposed feature takes more advantages for computational cost by calculating training time and prediction speed. We compared the proposed feature with the modern color texture features to highlight its advantage. The experiment in 10-Fold cross validation shows the robustness of our proposed feature in training and testing. In future, the proposed feature with plant disease classification system will be applied in agriculture domain to support economy and food productivity for farmers.

## References


[1] Camargo, A., and J. S. Smith. "An image-processing based algorithm to automatically identify plant disease visual symptoms." Biosystems engineering 102.1 (2009): pp.9-21.

[2] Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection." Frontiers in plant science 7 (2016): 1419.

[3] Wang, Guan, Yu Sun, and Jianxin Wang. "Automatic image-based plant disease severity estimation using deep learning." Computational intelligence and neuroscience 2017 (2017). pp. 28-46.

[4] Rangarajan, Aravind Krishnaswamy, Raja Purushothaman, and Aniirudh Ramesh. "Tomato crop disease classification using pre-trained deep learning algorithm." Procedia computer science 133 (2018): 1040-1047.

[5] Neumann, Marion, Lisa Hallau, Benjamin Klatt, Kristian Kersting, and Christian Bauckhage. "Erosion band features for cell phone image based plant disease classification." In Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 3315-3320. IEEE, 2014.

[6] Pujari, Jagadeesh D., Rajesh Yakkundimath, and Abdulmunaf S. Byadgi. "Image processing based detection of fungal diseases in plants." Procedia Computer Science 46 (2015): 1802-1808.

[7] Chouhan, Siddharth Singh, et al. "Bacterial foraging optimization based Radial Basis Function Neural Network (BRBFNN) for identification and classification of plant leaf diseases: An automatic approach towards Plant Pathology." IEEE Access 6 (2018): 8852-8863.

[8] Ferentinos, Konstantinos P. "Deep learning models for plant disease detection and diagnosis." Computers and Electronics in Agriculture 145 (2018): 311-318.

[9] López, Fernando, José Miguel Valiente, Ramón Baldrich, and María Vanrell. "Fast surface grading using color statistics in the CIE Lab space." In *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 666-673. Springer, Berlin, Heidelberg, 2005.

[10] Vertan, Constantin, and Nozha Boujemaa. "Color texture classification by normalized color space representation." In *Pattern Recognition, 2000. Proceedings. 15th International Conference On*, vol. 3, pp. 580-583. IEEE, 2000.

[11] Hanbury, Allan, Umasankar Kandaswamy, and Donald Adjeroh. "Illumination-invariant morphological texture classification." *Mathematical Morphology: 40 Years On* (2005): 377-386.