Texture histogram features for tea leaf identification using visible digital camera

B H Iswanto* and A Alma
Complex Systems Research Group, Department of Physics, FMIPA, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta 13220, Indonesia

*bhi@unj.ac.id

Abstract. This paper presents our study on the statistical texture histogram features to identify fresh tea leaves using a visible digital camera. For this purpose, the tea leaves were shouted every three days using the camera with 8 different orientations, a multiple of 45 degrees. Features of the images were extracted using the method to collect the feature dataset. Therefore, Principal Component Analysis (PCA), LBGU-EM clustering method, and Fisher’s Linear Discriminant Analysis (LDA) were applied to analyze the tea leaves based on the dataset. Experimental results using 320 image samples of four different categories show that the proposed method generated the image features that can significantly distinguish the fresh tea leaves categories.

1. Introduction

Tea (Camellia synesis) is an important commodity in the world whose quality is determined by several parameters, such as variety, environment, nutrition, and storage [1]. Plantation areas that generally have high rainfall are very conducive to the reproduction of various plant diseases so that it can have an impact on the decline in production and quality of tea. Therefore, intensive monitoring needs to be done to prevent and anticipate problems during growth. Usually, monitoring of the growth of tea is done by an expert by observing the physiology of tea leaves. However, the number of samples that can be observed per day is very limited, the accuracy is low, and often inconsistent [2], so it is not effective for large tea plantations area.

On the other hand, digital image processing technology has been developing and applied intensively to solve various agricultural problems, including recognizing and classifying plants leaves. This identification process includes a pre-processing stage, feature selection, and extraction, followed by identification and classification using a selection of proper classification techniques. Generally, image feature extraction is conducted by two different approaches. First, feature extraction is extracted based on image texture and colour characteristics; and the others based on thresholding and clustering techniques. In the literature, we can be found some applications, e.g. for the recognition of tea categories [3,4], detection of leaf disease [5], identification of leaf infection [6], and lesion characteristics [7]. However, there are still many problems remaining, among others the process is inefficient because it involves more computations and increases the processing time, especially in the feature extraction stage.
This paper reports our experimental study on statistical histogram features to recognize the fresh tea leaves using a visible digital camera. The feature data with different categories are analysed using the principal component analysis (PCA), cluster analysis, and linear discriminant analysis (LDA).

2. Methodology

2.1. Image acquisition
Our research used Canon 4000D camera to photograph many tea leaves from Gunung Mas Tea Plantation located at Cisarua Bogor, Jawa Barat Province, Indonesia. The camera placed in a special box equipped with two lamps and a sample holder. To analyse the growth phase of tea leaves, image samples were collected periodically every three days for one month. Shooting on a tea leaf sample is carried out eight times in different orientations, multiples of 45 degrees.

2.2. Pre-processing
The collected images are then pre-processed before further processing. Pre-processing aims to improve image data and reduce unwanted distortion or enhances some image features [8]. Our proposed method starts with the conversion of the RGB image into a single component grayscale image to reduce the computation cost.

2.3. Features selection and extraction
In image recognition, feature extraction is an essential step in the construction of any pattern classification and aims to extract the relevant features that characterize each class or category [9]. In this process, informative features are extracted from objects to the feature data formers. In this study, we investigate the statistical texture features to recognize the tea leaves. These features provide information about the properties of the image intensity, such as contrast, uniformity, flatness, and smoothness. The statistical features, such as mean, variance, skewness, kurtosis, energy, and entropy, dissimilarity, contrast, correlation, homogeneity, and autocorrelation can be computed using the texture histograms extracted from images [10]. Each feature has a specific measure of image texture. Practically, the image histograms are computed first. Then, the image features are extracted by using the probability distribution of pixel intensity levels in the histogram bins of the histogram [11]. Let P(b) is the probability distribution of bin b in the histogram N levels and f_b is the frequencies. The statistical measures of histogram like mean ($\mu$), variance ($\sigma^2$), skewness (S), kurtosis (K), energy (E), entropy (H), dissimilarity (D), contrast (C), correlation (Co), homogeneity (IDM), and autocorrelation (Ac) are defined as in Table 1.

| Table 1. The statistical measures of the histogram. |
|-----------------------------------------------|
| $\mu = \sum_{b=1}^{N} b \cdot P(b)$ | $Co = \sum_{b=1}^{N} \frac{b \cdot P(b) - \mu}{\sigma}$ | $H = -\sum_{b=1}^{N} P(b) \log_2 P(b)$ |
| $\sigma^2 = \sum_{b=1}^{N} (f_b - \mu)^2 P(b)$ | $Ac = \sum_{b=1}^{N} \frac{P(b + x, b + y)}{P(b)}$ | $C = \sum_{b=1}^{N} b^2 \cdot P(b)$ |
| $S = \frac{1}{\sigma^3} \sum_{b=1}^{N} (f_b - \mu)^3 P(b)$ | $K = \frac{1}{\sigma^4} \sum_{b=1}^{N} (f_b - \mu)^4 P(b)$ | $IDM = \sum_{b=1}^{N} \frac{P(b)}{1 + b^2}$ |
| $D = \sum_{b=1}^{N} P(b)$ | $E = \sum_{b=1}^{N} P(b)^2$ | |

2
The informative features should contain the information required to distinguish between categories, be insensitive to irrelevant variability in the input, and also be limited in number to permit efficient computation of discriminant functions as quoted by Lippman [12]. To obtain the informative features, most representative techniques frequently used are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [13]. PCA is unsupervised clustering, while LDA is a supervised classification technique. Also, cluster analysis is needed to find information about the structure of the data [14]. In this study, the LBGU-EM algorithm [15] is implemented. The algorithm is a combination of the EM algorithm and a vector quantization method called LBG-U so more consistent and robust from initialization problems.

3. Results and discussion
In this experimental study, 320 image samples of tea leaves have been collected, or 80 image samples for each category, to be analysed. These images consist of four categories, each category represents a different growth phase of a tea leaf. We labelled the sample data as Category1, Category2, Category3, and Category4 according to their growing age: 3, 6, 9, and 12 days respectively. The pre-processing stage was conducted to the image samples to remove background and convert to the single-channel greyscale images. Then, the features were extracted from the image samples by using equations as defined in Table 1. The average features of different leaf categories are as shown in Table 2. From Table 2 it is seen significantly different values of features for each category, except for the correlation, homogeneity, and autocorrelation.

| Features     | Category1 | Category2 | Category3 | Category4 |
|--------------|-----------|-----------|-----------|-----------|
| Mean         | 0.996878  | 0.970357  | 0.934570  | 0.843833  |
| Entropy      | 0.347829  | 0.619812  | 1.144420  | 2.429281  |
| Variance     | 0.007901  | 0.017901  | 0.037961  | 0.077413  |
| Skewness     | -6.435217 | -4.504282 | -2.823397 | -1.293546 |
| Kurtosis     | 38.446939 | 18.837919 | 6.148390  | -0.166759 |
| Energy       | 0.947934  | 0.906199  | 0.812500  | 0.587936  |
| Dissimilarity| 0.001884  | 0.004290  | 0.006483  | 0.016585  |
| Contrast     | 0.003141  | 0.005100  | 0.007690  | 0.015236  |
| Correlation  | 0.996936  | 0.997411  | 0.998207  | 0.998010  |
| Homogeneity  | 0.998845  | 0.997944  | 0.996810  | 0.991669  |
| Autocorrelation| 62.803724| 61.596258 | 58.737442| 51.296292 |

Furthermore, the Principal Component Analysis (PCA) was used to analyse the structure of the feature data. PCA generates the principal components, eigenvectors, and eigenvalues. For the features data, the first principal component (PC1) and second principal component (PC2) account for 88.90% and 7.40% of the total variation in the data respectively, or account for 96.30% of the total variation in the data. For data visualization purposes, the features are transformed to the principal components as scores then plotted on the PC1 vs PC2 space. PCA also generates loadings. Figure 1 shows the biplot of both scores and loadings together. It is seen that on the PC1, variance, entropy, contrast, dissimilarity, skewness, and correlation have large positive loadings, while the others have large negative loadings. Therefore, some features that have similar eigenvector and eigenvalue can be considered to be reduced.

Clusters of the objects have been analysed by using the LBGU-EM algorithm. According to the categories of tea understudied, the experiments were performed by using four-component densities
with the covariance matrices. In pre-processing, we normalized the variables of the data. The clustering results are presented in Table 3. From the table, there are a significant between cluster centroids in term of distances between cluster centroids and the within-cluster sum of squares error (SSE). These measures show that the statistical features can distinguish the tea leaves categories.

![Biplot of the statistical feature samples in PC1 and PC2 feature space.](image)

**Figure 1.** Biplot of the statistical feature samples in PC1 and PC2 feature space.

| Cluster1 | Cluster2 | Cluster3 | Cluster4 | Number of observation | Within cluster SSE |
|----------|----------|----------|----------|-----------------------|-------------------|
| Cluster1 | 0.0000   | 4.3489   | 2.0730   | 8.3257                | 80                |
| Cluster2 | 4.3489   | 0.0000   | 2.6271   | 4.7326                | 80                |
| Cluster3 | 2.0730   | 2.3805   | 0.0000   | 6.6642                | 80                |
| Cluster4 | 8.3257   | 4.7326   | 6.6642   | 0.0000                | 80                |

Furthermore, to investigate how the features contribute to group separation, we used the Linear Discriminant Analysis (LDA) method under the assumption that the covariance matrices are similar for all groups in the feature data. In the experimental study, the proposed method reached 100% accuracy, as shown in Table 4.
Table 4. Classification of feature data using linear discriminant analysis.

| True Group | Group1 | Group2 | Group3 | Group4 |
|------------|--------|--------|--------|--------|
| Group1     | 80     | 0      | 0      | 0      |
| Group2     | 0      | 80     | 0      | 0      |
| Group3     | 0      | 0      | 80     | 0      |
| Group4     | 0      | 0      | 0      | 80     |
| Total      | 80     | 80     | 80     | 80     |
| Correct    | 80     | 80     | 80     | 80     |
| Proportion | 100%   | 100%   | 100%   | 100%   |

Summary: Ntotal = 320, Ncorrect = 320, proportional correct = 100.0%

4. Conclusion

We have demonstrated that the statistical texture histogram can provide powerful features for fresh tea leaves recognition using a visible camera. Analysis results using the principal component analysis, cluster analysis, and discriminant analysis show high separability of image examples in the feature space and proportion correct until 100%. The methods can also identify some outliers and irrelevant features so that the number of features can be reduced. Future research can be directed at developing an identifier system using machine learning techniques by making more data acquisition. In addition, a study on this method is also needed to obtain a robust identification system and does not depend on lighting conditions.

References

[1] Karwowska K, Skotnicka M, and Smiechowska M 2019 Tea production and its forecasts, and the possibility of tea cultivation in the context of environmental requirements in China Problems of World Agriculture 19 1
[2] Dhingra G, Kumar V and Joshi H D 2018 Study of digital image processing techniques for leaf disease detection and classification Multimed Tools Application 77 19951–20000
[3] Wang S, Liu A, Philips P and Du S 2017 Tea category identification using computer vision and generalized eigenvalue proximal SVM Fundamenta Informaticae 151(1-4) 325-339
[4] Diniz P H G D, Dantas H V, Melo K D T, Barbosa M F, Harding D P, Nascimento E C L, Pistonesi M F, Bandb B S F, and Araújo M C U 2012 Using a simple digital camera and SPA-LDA modeling to screen teas Analytical Methods 4(9) 2648-2652
[5] Yang N, Yuan M, Wang P, Zhang R, Sun J and Mao H 2019 Tea diseases detection based on fast infrared thermal image processing technology Journal of the Science of Food and Agriculture 99 7
[6] Molina J F, Gil R, Bojacá C, Gómez F and Franco H 2014 Automatic detection of early blight infection on tomato crops using a color based classification strategy XIX Symposium on Image, Signal Processing and Artificial Vision 1-5
[7] Tucker C C and Chakraborty S 1997 Quantitative assessment of lesion characteristics and disease severity using digital image processing J Phytopathology 145(7) 273–278.
[8] Anjum M A and Javed M Y 2007 Multiresolution and varying expressions analysis of face images for recognition Information Technology Journal 6(1) 57-65
[9] Kumar G and Bhatia P K 2014 A detailed review of feature extraction in image processing systems Fourth International Conference on Advanced Computing & Communication Technologies 5-12
[10] Gonzalez R C and Woods R E 2017 Digital Image Processing 4th Ed (Pearson)
[11] Malik F and Baharudin B 2013 The statistical quantized histogram texture features analysis for image retrieval based on median and laplacian filters in the dct domain The International Arab Journal of Inf. Tech. 10(6)
[12] Pu X, Yi Z, Wu Y 2006 Recognizing partially damaged facial images by subspace auto-associative memories Advances in Neural Networks (Lecture Notes in Computer Science vol 3972) ed Wang J, Yi Z, Zurada J M, Lu BL, and Yin H (Berlin: Springer Heidelberg)
[13] Pandey P K, Yaduvir S and Sweta T 2011 Image processing using principal component analysis International Journal of Computer Applications 15(4) 37-40
[14] Jondya G and Iswanto B H 2017 Indonesian’s traditional music clustering based on audio features Procedia computer science 116 174-181.
[15] Iswanto B H and Fritzke B 2003 LBGU-EM algorithm for mixture density estimation Between Data Science and Applied Data Analysis (Studies in Classification, Data Analysis, and Knowledge Organization) ed Schader M, Gaul W and Vichi M (Berlin: Springer) 252-260