Revolutionize Date Fruit Classification using Optimized K-Nearest Neighbor

Khaidir Mauladan[1], **Wikky Fawwaz Al Maki**[2]

[1,2]School of Computing, Telkom University, Bandung
[1]khaidir@students.telkomuniversity.ac.id, [2]wikkyfawwaz@telkomuniversity.ac.id

Abstrak
Berbagai varian buah kurma yang menyebar di seluruh dunia memiliki kompleksitas dan karakteristik yang unik seperti warna, rasa, bentuk, dan tekstur. Studi ini mempertimbangkan fitur-fitur yang rumit dan beragam dari berbagai spesies buah kurma. Ciri-ciri unik ini dapat menimbulkan tantangan dalam menentukan dan membedakan antara berbagai jenis buah kurma yang ada. Klasifikasi buah-buahan ini dapat menjadi sangat sulit karena perbedaan-perbedaan halus dalam fitur yang telah disebutkan yang ada di antara spesies yang berbeda. Untuk mengatasi masalah ini, kategorisasi otomatis buah kurma telah muncul sebagai hasil dari pengembangan pembelajaran mesin dan visi komputer. Studi ini mengusulkan skema klasifikasi lima kategori untuk buah kurma. Untuk menghasilkan data pelatihan berkualitas tinggi untuk pengklasifikasi K-Nearest Neighbor (KNN), ekstraksi fitur dilakukan pada gambar buah kurma menggunakan momen warna, circularity, dan deskriptor HOG masing-masing. Analisis Komponen Utama (PCA) digunakan sebagai pendekatan reduksi dimensionalitas untuk meningkatkan model KNN yang diusulkan. Untuk menentukan fitur terbaik yang ditentukan berdasarkan pengaruhnya pada kinerja model, Binary Particle Swarm Optimization (BPSO) digunakan sebagai pendekatan pemilihan fitur. Model klasifikasi mencapai tingkat akurasi 93,08%, yang merupakan peningkatan yang cukup besar sebesar 10,77% dibandingkan dengan model KNN dasar.

Kata Kunci: Binary Particle Swarm Optimization, Circularity, Color Moments, Histogram of Oriented Gradients, K-Nearest Neighbor, Principal Component Analysis

Abstract
Different variations of date fruit spreading around the globe possess their unique complexity and distinctive characteristics such as color, taste, shape, and texture. This study took into account the intricate and diverse features of various species of date fruit. These unique traits can present challenges in accurately defining and distinguishing between the numerous types of date fruit that exist. The classification of these fruits can be particularly difficult due to the subtle differences in the aforementioned features that exist between the different species. To solve this issue, automatic date fruit categorization has emerged as a result of the development of machine learning and computer vision. This study proposes a five-category classification scheme for date fruit. To generate high-quality training data for the K-Nearest Neighbor (KNN) classifier, feature extraction was done on the date fruit images using color moments, circularity, and HOG descriptors respectively. Principal Component Analysis (PCA) was employed as a dimensionality reduction approach to improving the suggested KNN model. To determine the best features determined by their influence on the model's performance, Binary Particle Swarm Optimization (BPSO) was used as a feature selection approach. The classification model attained an accuracy rate of 93.08%, which is a considerable increase of 10.77% over the basic KNN model.

Keywords: Binary Particle Swarm Optimization, Circularity, Color Moments, Histogram of Oriented Gradients, K-Nearest Neighbor, Principal Component Analysis

1. Introduction

Background

Date fruits are a good source of nourishment and are commonly grown in hot and dry areas across the world [22]. Accurate and dependable fruit categorization is critical for improving the efficiency of date fruit production and quality management. Date fruit has an extensive cultivation history stretching back to previous civilizations. Date fruit variations abound over the world, including but not limited to Galaxy, Sugaey, Shaishe, Ajwa, and Nabtat Ali, each with its own particular texture, flavor, color, and shape. However, there are various obstacles to overcome while classifying date fruits. One difficulty is the visual variety of date fruits, since different varieties may have slight changes in color, shape, and appearance. Automatic date fruit categorization has emerged as a result of the development of machine learning and computer vision, with the potential to use these technologies for robotic harvesting [3]. Automatic categorization not only benefits in classifying date fruits, but it can also benefit other sectors as well to ensure the consistency of the target quality [5]. A study by [1] examined four strategies with the Support Vector Machine (SVM) having the greatest accuracy of all four techniques employed. Another study by [17] utilized a Deep CNN to develop a date fruit classification model for automated date fruit sorting. The study by [17] could classify three distinct kinds of date fruits with great accuracy above 95%. The developed model in the mentioned study was tested and evaluated on three different types of date fruit. Both studies has not demonstrated that it functions effectively with extra classes. More research is needed to fully comprehend its possibilities as well as its limitations.

In image classification, the data are labeled by their corresponding classes. Therefore, image classification is regarded as a supervised learning problem. Supervised learning image classification implies predicting a class label based on an image's distinct features and patterns. This study's approach to solving the supervised learning problem is by employing K-Nearest Neighbor (KNN). However, two studies by [18] and [16] compared their proposed image classification technique with barebone KNN resulted in inadequate performance on barebone KNN. Both KNN models do not reach 75% accuracy performance. This occurrence happened because both study did not extract influential features of an image for KNN to utilize.

The issue in image classification is determining which aspects of a picture are meaningful and may have a substantial influence on the performance of the model being employed. Color, texture, shape, and size, among other factors, might all have a part in influencing the accuracy of the categorization. To obtain optimal performance in image classification tasks, it is critical to carefully analyze the selection and extraction of these characteristics. Additionally, an image segmentation procedure is beneficial for feature extraction as it can separate the target of interest from its surroundings making feature extraction more precise to the target [9]. A study by [21] demonstrates color and shape features perform better than using only one of them in classifying flowers. Another study by [14] indicated that color and shape attributes may be used to distinguish fruits other than apples and can improve the model's performance. Furthermore, it is stated that color features can be used to distinguish fruits with diverse hues, such as date fruits. Additionally, date fruits have distinct structures of creases present on their surface. Due to the irregular nature of the creases that form the surface characteristics of date fruits, it is necessary to utilize image descriptors to capture the texture patterns present in the image as the Pyramid Histogram of Orientation Gradients (PHOG) technique was successfully applied to classify three large cat species—Cheetah, Jaguar, and Tiger—based on their fur pattern, with an accuracy of 91.07% employing SVM as the classifier method in a study by [16]. In this study, an optimization technique for feature selection is proposed to improve the model's performance.

Optimization algorithms are methods for determining the best solution to a problem involving an objective function. Population-based optimization approaches are esteemed to be efficient and have minimal computing labor. One population-based optimization approach is Particle Swarm Optimization (PSO). However, feature selection using PSO is conceptually unsuitable as PSO generates a continuous value that does not correlate with the concept of feature selection. To utilize PSO as a feature selection technique, Kenned and Eberhart proposed a variation of PSO, known as Binary PSO in 1995 [13]. Several studies by [15], [6], and [7] implemented a PSO-based optimization technique for feature selection and successfully improved their classifier method with a considerable increase in accuracy.

Problem Statement

The problem statement for this research is to develop an effective image classification model for date fruit that takes into account the irregular surface characteristics and texture patterns of the fruit, and can accurately classify multiple varieties. Despite previous studies utilizing image descriptors and KNN models, their proposed techniques have shown inadequate results with accuracies below 75%. Additionally, the current model that was developed and tested in [2] has not been proven to perform well with additional classes. Therefore, there is a need for further research to fully understand the capabilities and potential limitations of the model, and to explore alternative optimization approaches such as Binary Particle Swarm Optimization to improve the performance of KNN for the date fruit image classification technique.

Objective

This study aims to construct a profound classification model that is capable of distinguishing distinctive kinds of date fruit by taking into account the irregular surface characteristics and texture patterns of the fruit. The proposed technique should be able to improve upon the inadequate results shown in previous studies, which had accuracies below 75%. Additionally, the objective is to explore alternative optimization approaches such as Binary Particle Swarm Optimization to improve the performance of the image classification technique for date fruit and to fully understand the capabilities and potential limitations of the model.

Report Structure

The structure of this study is summarized as follows: Section II describes the proposed research approach, which includes image preprocessing, image feature extraction, and dimensionality reduction. The retrieved features are then put into BPSO to optimize influential features, which are subsequently used for classification and evaluation. Section III explains and analyzes the study's findings. Section IV summarizes the study's findings.

2. Literature Review

2.1 Color Moments

Color moments is an efficient and commonly used approach for representing an image's color properties such as the measure of brightness and intensity notably in image processing [20]. As it gives a concise description of the color content and is immune to variations in the image, color moments is a recommended solution for tasks such as object identification and image categorization. The RGB and HSV color spaces are generally used to determine color moments. Three color moments are calculated comprises of the color channel's mean, skewness, and variance to represent the color's distribution [23]. The mean depicts the average intensity of the colors in the picture, whereas the variance quantifies the color dispersion around the mean. However, skewness describes the asymmetry of the color distribution. The mathematical model of the three color moments is as (1), (2), and (3) respectively [23].

$$E_i = \frac{1}{N}\sum_{j=1}^{N} p_{i,j} \tag{1}$$

$$\gamma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(p_{i,j} - E_i)^2} \tag{2}$$

$$\sigma_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(p_{i,j} - E_i)^3} \tag{3}$$

where N denotes the total number of pixels of the image and $p_{i,j}$ represents the current pixel's coordinate.

2.2 Circularity

Circularity is the degree to which a shape deviates from a complete circle. Circularity is a metric that measures how closely an object boundary resembles a circle in a range of inclusion (0, 1), with a shape rated as 1 if and only if it is a perfect circle. The circularity formula is represented in (4) [24].

$$C = 4\pi A/perimeter^2 \tag{4}$$

where A and $perimeter$ denote the area and perimeter of an object in the image.

2.3 Histogram of Oriented Gradients (HOG)

HOG is a feature descriptor that keeps track of how often gradient orientation appears in a detection window [25]. HOG separates the image into numerous cells, which are then grouped into blocks. HOG calculates the gradient magnitude and orientation for every pixel within a block using (5) and (6) [2].

$$G = \sqrt{g_x^2 + g_y^2} \tag{5}$$

$$\theta = \arctan(\frac{g_x}{g_y}) \tag{6}$$

where g_x and g_y are gradients in the horizontal and vertical direction computed after one dimension filtering [2]. Following the gradient computation, a histogram for each block in the separated image is computed. Each pixel's gradient magnitude is separated into orientation bins depending on the angle, and then neighboring blocks are aggregated and normalized using L_2-normalization to generate a histogram description.