

WEED DETECTION USING K-NEAREST NEIGHBOUR ALGORITHM

ARISHAH LIM

A report submitted
in fulfillment of the requirements for the degree of
Bachelor of Mechanical Engineering




UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2021

DECLARATION

I declare that this project report entitled “Weed Detection Using K-Nearest Neighbour Algorithm” is the result of my own work except as cited in the references.

	Signature	:
	Name	:	Arishah Lim
	Date	:

اونيورسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

APPROVAL

I hereby declare that I have read this project report and in my opinion this report is sufficient in terms of scope and quality for the award of the degree of Bachelor of Mechanical Engineering.

	Signature	: 
	Supervisor's Name	: Mr. Zairulazha bin Zainal
	Date	: 18-5-2021

اونيورسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DEDICATION

To my beloved mother, father and siblings for the endless support



ABSTRACT

Weed management is one of the important aspects in agriculture sector but traditional weeding method leads to unnoticed growth of weeds which cause loss in production. Early weed detection for plantation health monitoring helps to improve the overall crops yielding. This can be achieved by approaching technology of supervised machine learning for weed detection. The objectives of this project are to define the procedures and methodology in using k-Nearest Neighbours for weed detection, to analyse the performance of the algorithm built and to study the relationship between value of parameter k and rate of accuracy. MATLAB coding and applications were used to develop the weed detection model for classification of weeded and non-weeded areas. Data acquisition was done through obtaining images of areas with soil, average weeds and abundant weeds. The Bag of Visual Words (BoVW) technique was then utilized for extracting features from the image dataset with grid method used for selection of feature key points. The nearest neighbour classifier was trained with the features extracted on training set of images and validated on test set. The analysis of performance was done based on confusion matrix chart which shows the accuracy, precision, recall and F-measure of the trained model. From the simulation, Fine k-Nearest Neighbour with one nearest neighbour had provide the highest accuracy which is 0.98 for training and 0.85 for validation of classifier. The precision and recall for training were 1.00 and 0.97 whereas for validation were 0.77 and 1.00. By changing the value of k for the k-Nearest Neighbour classifier, result showed that the accuracy of model decreases with the increasing of k. The validation accuracy was found to be slightly lower than the training accuracy, which could be due to insufficient work in hyperparameter tuning. Overall, k-Nearest Neighbour algorithm has shown potential of efficiency and reliability in weed detection. Future work should focus on the optimization of the model in order to further improve the performance.

ABSTRAK

Pengurusan rumpai merupakan salah satu aspek yang penting dalam sektor pertanian, tetapi kaedah merumput yang biasa membawa kepada pengabaian dalam pertumbuhan rumpai. Kejadian ini menyebabkan kerugian dalam pengeluaran hasil tanaman. Pengesanan rumpai pada tahap awal untuk pemantauan perladangan boleh membantu dalam penambahbaikan hasil tanaman. Ini boleh dicapai dengan menggunakan teknologi pembelajaran mesin terarah untuk pengesanan rumpai. Objektif projek ini adalah untuk menentukan prosedur dan kaedah dalam menggunakan algoritma k -jiran terdekat (k -NN) untuk pengelasan kawasan berumpai. Prestasi model algoritma yang dibina akan dianalisa dan hubungan antara nilai ' k ' dengan kadar ketepatan juga akan dicerap. Pengekodan dan aplikasi MATLAB telah digunakan untuk membangunkan model pengesanan rumpai bagi klasifikasi antara kawasan tanpa rumpai dan berumpai. Pangkalan data imej telah dibina berdasarkan kawasan tanpa rumpai, komposisi rumpai yang sederhana dan komposisi rumpai yang tinggi. Teknik beg perkataan visual (BoVW) merupakan pendekatan untuk pengekstrakan ciri daripada imej dengan kaedah grid bagi pemilihan lokasi ciri tersebut. Menggunakan ciri yang diekstrak, model k -NN dilatih dengan set imej latihan dan disahkan dengan set imej ujian. Analisis prestasi model dijalankan dengan menggunakan kaedah matriks kekalutan yang menunjukkan ketepatan, kepersisan, kepekaan dan nilai ukuran- F . Berdasarkan keputusan simulasi, k -NN halus didapati mempunyai ketepatan yang tinggi dalam latihan model (0.98) dan pengesanan model (0.85). Nilai kepersisan dan kepekaan bagi proses latihan adalah 1.00 dan 0.97, manakala bagi pengesanan adalah 0.77 dan 1.00. Keputusan juga menunjukkan kadar ketepatan model menurun apabila nilai ' k ' meningkat. Kemungkinan kadar ketepatan untuk pengesanan model lebih kurang berbanding dengan latihan model adalah kerana kekurangan penalaan hiperparameter. Model k -NN telah menunjukkan potensi yang baik dalam pengelasan kawasan berumpai secara keseluruhan. Penyelidikan pada masa hadapan boleh memberi perhatian atas pengoptimuman model k -NN bagi tujuan menambahbaik prestasi keseluruhan.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my supervisor En. Zairulazha bin Zainal for all the kindness in providing guidance and endless support towards the completion of this final year project.

I would like to show gratitude to my examiners, Dr. Asriana binti Ibrahim and Dr. Rainah binti Ismail for giving me useful advice and suggestions during the seminar. I would also like to thank Universiti Teknikal Malaysia Melaka (UTeM) and Faculty of Mechanical Engineering for giving me this opportunity to participate and gain experience through this final year project.

Last but not least, special thanks to my peers, my mother, father and siblings for their constant support. Their endless support is encouragement and motivation to myself for giving my best in completing this project.

TABLE OF CONTENTS

CHAPTER	CONTENT	PAGE
	DECLARATION	
	APPROVAL	
	DEDICATION	
	ABSTRACT	i
	ABSTRAK	ii
	ACKNOWLEDGEMENTS	iii
	TABLE OF CONTENTS	iv
	LIST OF TABLES	vi
	LIST OF FIGURES	vii
	LIST OF ABBREVIATIONS	ix
	LIST OF SYMBOLS	x
CHAPTER 1	INTRODUCTION	1
	1.1 Background	1
	1.2 Problem Statement	3
	1.3 Objective	3
	1.4 Scope of Project	4
CHAPTER 2	LITERATURE REVIEW	5
	2.1 Weed Detection	5
	2.2 Machine Learning Algorithms	6
	2.3 Classification Techniques for Weed Detection	8
	2.4 K-Nearest Neighbour Algorithm	10
	2.5 Image Processing	12
	2.6 Confusion Matrix	14
CHAPTER 3	METHODOLOGY	16
	3.1 Introduction	16
	3.2 General Process	16
	3.3 Data Collection	18
	3.4 Image Processing	20
	3.4.1 Image Database	20
	3.4.2 Feature Extraction	21
	3.5 Weed Detection Classifier	22
	3.5.1 Training of Classifier	22
	3.5.2 Testing of Classifier	24

3.6	Method of Analysis	25
CHAPTER 4	RESULTS AND ANALYSIS	27
4.1	Introduction	27
4.2	Data Processing	28
4.3	Feature Extraction	30
4.4	Training of k-NN Classifier	32
4.5	Testing of k-NN Classifier	33
	4.5.1 Visualizing Weed Detection Result	34
4.6	Analysis on Performance of Simulation	36
	4.6.1 Performance of Feature Extraction	36
	4.6.2 Performance of Training of Classifier	37
	4.6.3 Performance of Validation of Classifier	40
CHAPTER 5	CONCLUSION	42
5.1	Conclusion	42
5.2	Recommendations	43
	REFERENCE	44
	APPENDICES	47
	Appendix A	47
	Appendix B	48
	Appendix C	50



LIST OF TABLES

TABLE	TITLE	PAGE
2.1	Types of classification techniques	9
2.2	Confusion matrix	14
3.1	Types of nearest neighbour classifiers	24
3.2	Confusion matrix of weed detection	26
4.1	The count of images from each category	28
4.2	Bag of features extraction details	30
4.3	Feature data for first six vocabularies	32
4.4	Comparison of key point extraction methods	37
4.5	Results of accuracy for different trained model	37

LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1	Supervised machine learning process	6
2.2	Example of scatter plot of k-NN	11
2.3	Feature point selection using detector and grid method	13
3.1	Process flow of project	15
3.2	Images of soil	18
3.3	Images of area with average weeds	19
3.4	Images of area with abundant weeds	19
3.5	MATLAB coding for creating image database	20
3.6	Process flow of feature extraction using bag of visual words	21
3.7	MATLAB coding for image feature extraction	22
3.8	Preparation of data in Classification Learner	23
3.9	MATLAB coding for validation of classifier	25
4.1	Sample images from non-weeded category	29
4.2	Sample images from weeded category	29
4.3	Histogram for non-weeded image	31
4.4	Histogram for weeded image	31
4.5	Details of the trained k-NN model	32
4.6	Scatter plot of trained Medium KNN model	33
4.7	Accuracy of test result	34
4.8	Coding to visualize weed detection	34

4.9	Correct prediction of non-weeded area	35
4.10	Correct prediction of weeded area	35
4.11	False prediction of weeded area	36
4.12	Confusion matrix of trained Fine KNN model	38
4.13	Graph of accuracy against value of k	39
4.14	Confusion matrix of validation result	40



LIST OF ABBEREVATIONS

ANN	Artificial Neural Network
FP	False Positive
FN	False Negative
KNN	k-Nearest Neighbour
RF	Random Forest
RGB	Red, Green and Blue
SVM	Support Vector Machine
TP	True Positive
TN	True Negative



LIST OF SYMBOLS

d	-	Distance
x	-	Distance coordinate
y	-	Distance coordinate



CHAPTER 1

INTRODUCTION

1.1 Background

Agriculture sector is one of the important sectors in Malaysia that contributes to the economy of the country. According to Department of Statistics Malaysia, the agriculture sector has contributed RM 99.5 billion to the Gross Domestic Product (GDP) in 2018. Majority contribution of 65.8% were from plantation of crops which includes oil palm, rubber and paddy. As crops plantation holds a crucial position in the sector, it is essential for people to research deep into any challenges faced in the process and seek for solutions or improvements.

Weeds are long-term problem faced in plantation of crops which it could be harmless to the latter in beginning stages but turns into a big threat at the end. It was estimated that the yield loss in paddy fields in Malaysia due to grasses, broad-leaved weeds and sedges was 41%, 28% and 10% respectively (Karim et al. 2004). Undeniably, untreated weeds could cause big impact on production output which leads to major loss in crops earnings.

Weed is defined as a wild plant growing where it is not wanted especially among crops or garden plants. Weeds grow easily and often spread fast to nearby lawn due to their surviving abilities which most conditions are considered favourable to them. On same soil, weeds tend to outgrow crops in competing for resources like nutrients, water sunlight and space because the former is most likely already lying in the soil before crops are planted.

Hence, crops are unable to receive adequate nutrients which cause them become vulnerable to disease.

Although the growth of weeds is unavoidable, this situation is curbed by the process of weed control management. A variety of approaches that are employed in the sector today are hand weeding, mechanical weeding, slashing, burning, flooding, covering with organic or inorganic materials and using herbicides (Rutherford et al. 2011). The techniques mentioned are applied only when weeds are discovered by growers. Moreover, weed management in local vegetables and fruit farms usually relies on manual weeding or using herbicides. To reduce the risk of weeds competing with crops, the threat must be discovered as early as possible. Therefore, weed detection is important and should be introduced to restrict growth of weeds in early stage.

A weed classifier can differentiate between weeds and others and if it is combined with other technologies, it could assist plantations in other ways such as developing successful management plan for weed control. Some of the known classifiers are k-Nearest Neighbour (k-NN), Thresholding classifier and Support Vector Machine (SVM).

K-Nearest Neighbour is a simple classifying approach in the machine learning methods where the classification is attained through the identification of the adjacent associates to enquiry illustrations and then utilizes those associates for determining the group or class of the doubt (Khurana and Bawa, 2019). In simpler explanation, after learning features for two sets of data, k-NN classifies an unknown or new data based on the majority vote of the neighbouring data. 'k' is a value that need to be determined as it will affect the accuracy of the outcome. Before training a classifier, some of the important steps are data acquisition and data processing which the latter involved the extraction of image features to train the classifier.

1.2 Problem Statement

The existence of weeds in vegetables and fruits farms has affected the growth of the crops due to competition of sunlight, spaces, water and nutrients. Weeds are also shelter to pest and insects that could infect healthy crops with diseases. Practice of local farms using manual weeding technique is time and energy consuming as farmers had to constantly monitor the condition by themselves which leads to neglect in weed management. By introducing the use of weed detection using k-NN algorithm, not only time and energy can be saved, but also weeds will be detected right after growing out from soil thus avoiding the chance of becoming shelter to pests. This can contribute to better yield of crops and at the same time preserve the quality of the crops. Machine learning approach for weed detection can be a method of plantation health monitoring. In order to achieve better yield of crops, the classifier built should be reliable to use, such that the algorithm can detect weed accurately and precisely.

1.3 Objective

The objectives of this project are:

1. To define the procedures and methodology in using supervised learning of k-Nearest Neighbour algorithm to detect weeds.
2. To analyse the performance k-Nearest Neighbour algorithm in weed detection by means of accuracy and succession in the detection process.
3. To study the relationship between value of parameter k and rate of accuracy.

1.4 Scope of Project

The scopes of this project are:

1. Image processing and simulation of weed detection using k-NN classification will be carried out using MATLAB. Details of image processing will be discussed and included in this report.
2. In investigating the succession of weed detection using KNN, the parameters that should be recorded will include the number of true positive, true negative, false positive and false negative detection.
3. The effect of parameter k in detection accuracy will be studied by changing the value of k (e.g. 1-NN, 2-NN, ...) and observing the results of detection.



CHAPTER 2

LITERATURE REVIEW

2.1 Weed Detection

Introducing automation in weed detection can increase the efficiency of weed control in agricultural field to prevent yield loss and preserve the quality of crops. The methods are always related to computer vision, artificial intelligence, robotics, machine learning and more.

Wu et. al (2011) proposed a weed detection method using machine vision based on position and edge features. The authors started by segmenting the soil background from weeds and crops through thresholding on gray image. Crops and weeds were then differentiated by extracting the position and edge feature of crops because the arrangement of crops in the sampling location was fixed. The algorithm achieved accuracy rate between 92% and 95% and the authors discussed that the false detection is due to weakness in edge feature extraction and misjudgement of crops and weeds overlapping. There are limitations in the method mentioned because the algorithm is more suitable to be used in situation that position of crops were fixed and arranged.

A study by Hashim et.al (2019) shows research on using Convolution Neural Network (CNN) in weed detection. In this study, images were pre-processed using data augmentation through performing geometric transformation and change of colours, brightness and contrast. The train result of CNN in weed detection with data augmentation showed 95% accuracy whereas the test result showed 70.5% accuracy. However, the

authors managed to achieve a better output of test result (85.5%) when original image was used. CNN can be used for weeds detection, but the proposed method works more complicated than using supervised machine learning method.

Supervised machine learning is a more convenience method and many studies had focused on application of machine learning in weed detection as it is efficient in data handling other than performs well in solving classification problems. The pre-processed of images as well as choice of classifier requires more research to determine the most efficient method. This project focused on applying k-Nearest Neighbour in weed detection and the method of image processing will be discussed.

2.2 Machine Learning Algorithms

There are three categories under machine learning algorithms which are supervised machine learning, unsupervised machine learning and reinforcement machine learning. Supervised machine learning is about training the machine to learn sets of data before making predictions on other sets of test data whereas unsupervised machine learning is vice versa which no training on categorization or classification is given. This project focus on supervised machine learning in weed detection. Figure 2.1 shows the general process of supervised machine learning.

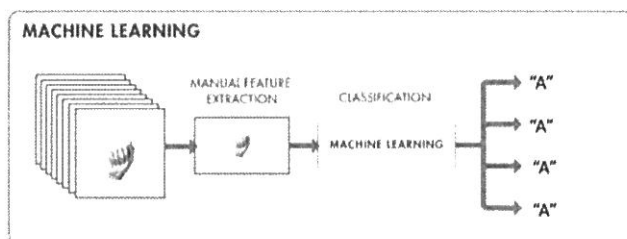


Figure 2.1: Supervised machine learning process (Murphy, 2020)

In a review journal written by Abas et. al (2020), the authors learned that the top ranked algorithms in supervised machine learning which are commonly used are Support Vector Machine, Artificial Neural Network, k-Nearest Neighbours, Linear Regression and others. The ranking reviewed the quality of the algorithms based on their characteristics like accuracy, strength, data size, extendibility and computational time in constructing the model. The authors also stated that machine learning algorithms are a part of artificial intelligence.

A study by Praveena et. al (2017) explained supervised machine learning with vector quantity as input object and supervisory signal as output value. The algorithms analyse the practice data and construct a modelling function in order to perform predictions on new examples. The authors listed out steps performed in supervised machine learning algorithms which are establishing type of training examples, converging training set, resolving input feature illustration, assimilating and executing the learning algorithm and lastly evaluating the accuracy of the function. The study concluded that each learning algorithm has different level of performance based on the problem types and probably requires configurations and adjustments on parameters to achieve optimal performance.

Muhammad & Zhu (2015) did survey on supervised machine learning algorithms and discussed on the issues in using this learning method. It is stated that better selection of features is essential to perform higher quality training on data set. Thus, an approach of “brute-force” is applied which is executing everything available without knowing whether the right features are isolated when the requirement is not met. This method caused noise and missing features value in data set that leads to requirement of additional pre-processing which is key function in supervised machine learning.

In the work of Dasgupta & Nath (2016), the authors mentioned that there are mainly two groups under supervised learning which are regression and classification. The output in regression is a real number or a whole vector of real number while the outcome of classification is a class label. The work also discussed on problems faced with supervised learning such as bias variance trade off, complexity of training data amount and function, input space dimensionality noise in input values, data integrity, data redundancy and presence of interactions and non-linearities. The difference between supervised learning and unsupervised learning is that the former will always produce the same outcome for the chosen input whereas the latter produces different outcomes on each run of the chosen input.

Gmyzin (2017) stated in his study that supervised machine learning is used in a wide field of applications including adaptive websites, natural language processing, healthcare and engineering. The different types of algorithms under supervised learning are logic based, perceptron based, statistical learning, instance-based learning, support vector machines and regression.

2.3 Classification Techniques for Weed Detection

Each and every classification technique works differently with different inputs and also affected by the parameters taken for training the algorithms. Table 2.1 shows several classification techniques under supervised machine learning algorithms. Most of these techniques are widely used in weed detection.

Table 2.1: Types of classification techniques

Algorithm Type	Techniques
Logic based algorithms	C4.5
Perceptron based techniques	Artificial Neural Network (ANN)
Statistical learning algorithms	Naïve Bayes classifiers
Instance based learning	k-Nearest Neighbour (k-NN)
Support Vector Machines	Support Vector Machines (SVM)
Regressions	Logistic regression

Weed detection is essential in sustainability in agriculture because weeds are often difficult to detect in time and separated with crops (Liakos et.al, 2018). Using classification techniques to discriminate weeds from crops has advantage of low cost and no side effects on the environment. These methods of weed detection encouraged more detailed studies in machine learning towards modern agricultural practices.

In study of Pulido Rojas et. al (2016), the authors did a comparative analysis among three classifiers which are Thresholding, k-NN and SVM for weed detection. Thresholding classifier in weed detection is based on area features whereas k-NN and SVM are based on texture patterns. The results of the study showed that k-NN classifier has the highest sensitivity as it has high ability to detect weeds correctly. On the other hand, SVM has highest specificity value while Thresholding had a result that it can correctly detect weeds but also has high number of false detections. However, the authors mentioned that the computing time for k-NN and SVM are relatively high compared to Thresholding which only requires 0.64s.

Yano et. al (2017) worked on selecting classifier for weed detection in sugarcane fields. The experimented classifiers were ANN, Random Forest (RF) and k-NN. The study involved sugarcane, soil and various weed species which are peppergrass, shoo-fly plant and signal grass. The authors found out that ANN classifier had the best performance in detecting most species, following by RF and weakest is k-NN which only good at detecting soil. Based on the results, it can be seen that the performance of classifiers is different when applied on different cases.

Each classifier has their own strength and results in varies outcomes based on different conditions. Dadashzadeh et. al (2020) chose to use ANN as method of weed detection because it has high computation speed and is capable in dealing with noisy input. The author achieved an average accuracy rate of 91% in weed classification. The study was further improved by using another algorithm, k-NN classifier to compare the results obtained. It was observed that k-NN generated more misclassification of rice as weeds in this case.

Therefore, the performance of the classifier should be evaluated from different aspects including the procedure of training, the accuracy of detection, the execution time and other factors. The algorithm used in this project is k-Nearest Neighbour algorithm which the performance of the classifier will be discussed later on.

2.4 K-Nearest Neighbour Algorithm

k-NN is an algorithm that classifies a test data to the nearest type of data that is proximity to it based on the training data. The value of 'k' represents the ungrouped data is classified by how many nearest neighbouring points. The classifier did it by calculating the distance between the mathematical value between these points. Some of the distance