Faculty of Information and Communication Technology

A HAZE REMOVAL TECHNIQUE FOR SATELLITE REMOTE SENSING DATA BASED ON SPECTRAL AND STATISTICAL METHODS

Nurul Iman Binti Saiful Bahari

Master of Science in Information and Communication Technology

2016

A HAZE REMOVAL TECHNIQUE FOR SATELLITE REMOTE SENSING DATA BASED ON SPECTRAL AND STATISTICAL METHODS

NURUL IMAN BINTI SAIFUL BAHARI

A thesis submitted
In fulfilment of the requirements for the degree of Master
in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

DECLARATION

I declare that this thesis entitled ―A Haze Removal Technique for Satellite Remote Sensing Data Based on Spectral and Statistical Methods" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :

Name : Nurul Iman Binti Saiful Bahari

Date : 19th May 2016

APPROVAL

I hereby declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality as a partial fulfilment of Master of Science in Information and Communication Technology.

Signature	:	
Supervisor Name	:	Dr. Asmala Bin Ahmad
Date	:	19^{th} May 2016

DEDICATION

I dedicate this work to my family.

My late father

Saiful Bahari Bin Abu Bakar

My late mother

Saedah Binti Abdul Rahman

My big brother

Loqman Hakim Bin Saiful Bahari

My younger sisters

Nurul Fatihah Binti Saiful Bahari

Sufiah Binti Saiful Bahari

I dedicate my love to you all always in this world and the Hereafter.

ABSTRACT

Haze originated from forest fire burning in Indonesia has become a problem for South-east Asian countries including Malaysia. Haze affects data recorded using satellite due to attenuation of solar radiation by haze constituents. This causes problems to remote sensing data users that require continuous data, particularly for land cover mapping. There are numbers of haze removal techniques but these techniques suffer from limitations since they are developed and designed best for particular regions, i.e. mid-latitude and high-latitude countries. Almost no haze removal techniques are developed and designed for countries within equatorial region where Malaysia is located. This study is meant to identify the effects of haze on remote sensing data, develop haze removal technique that is suitable for equatorial region, especially Malaysia and evaluate and test it. Initially, spectral and statistical analyses of simulated haze datasets are carried out to identify the effects of haze on remote sensing data. Land cover classification using support vector machine (SVM) is carried out in order to investigate the haze effects on different land covers. The outcomes of the analyses are used in designing and developing the haze removal technique. Haze radiances due to radiation attenuation are removed by making use of pseudo invariant features (PIFs) selected among reflective objects within the study area. Spatial filters are subsequently used to remove the remaining noise causes by haze variability. The technique is applied on simulated hazy dataset for performance evaluation and then tested on real hazy dataset. It is revealed that, the technique is able to remove haze and improve the data usage for visibility ranging from 6 to 12 km. Haze removal is not necessary for data with visibility more than 12 km because able to produce classification accuracy more than 85%, i.e. the acceptable accuracy. Nevertheless, for data with visibility less than 6 km, the technique is unable to improve the accuracy to the acceptable one due to the severe modification of spectral and statistical properties caused by haze.

ABSTRAK

Jerebu yang berasal dari pembakaran hutan di Indonesia telah menjadi masalah kepada negara - negara Asia Tenggara termasuk Malaysia. Jerebu memberi kesan kepada data yang direkod menggunakan satelit disebabkan oleh pelemahan radiasi suria oleh unsur-unsur jerebu. Ini menjadi masalah kepada pengguna data remote sensing yang memerlukan data yang berterusan, terutamanya untuk pemetaan litupan tanah. Telah wujud beberapa teknik penyingkiran jerebu tetapi teknik-teknik ini mengalami beberapa batasan kerana dibangunkan dan direka khas untuk kawasan-kawasan tertentu seperti negara-negara latitud pertengahan dan tinggi. Hampir tiada yang dibangunkan dan direka untuk negara di dalam kawasan khatulistiwa di mana Malaysia terletak. Kajian ini bertujuan untuk mengenalpasti kesan jerebu pada data remote sensing, membangunkan teknik penyingkiran jerebu yang sesuai untuk kawasan khatulistiwa, terutama Malaysia dan menilai dan mengujinya. Permulaannya, analisis spektrum dan statistik dilakukan pada satu set data simulasi jerebu untuk mengenalpasti kesan jerebu pada data remote sensing. Pengkelasan litupan tanah menggunakan 'support vector machine' (SVM) dilakukan untuk mengkaji kesan jerebu pada jenis litupan tanah yang berbeza. Hasil analisis seterusnya digunakan untuk mereka dan membangunkan teknik penyingkiran jerebu. Radian jerebu disebabkan oleh pengurangan radiasi disingkirkan dengan menggunakan 'pseudo invariant features (PIFs)' yang dipilih di antara objek reflektif di dalam kawasan kajian. Seterusnya, tapisan spatial dilakukan untuk menyingkirkan gangguan selebihnya yang disebabkan oleh kepelbagaian jerebu. Teknik ini diaplikasikan pada set data simulasi jerebu untuk menilai prestasinya dan kemudian diuji pada data jerebu sebenar. Telah didapati bahawa teknik ini mampu untuk menyingkir jerebu dan meningkatkan kebolehgunaan data untuk jarak ketampakan dari 6 ke 12 km. Penyingkiran jerebu adalah tidak diperlukan untuk data dengan jarak penampakan lebih dari 12 km kerana ia boleh menghasilkan ketepatan pengkelasan lebih dari 85%, iaitu ketepatan yang boleh diterimapakai. Walaubagaimanapun, untuk data dengan jarak penampakan kurang dari 6 km, teknik ini tidak mampu untuk meningkatkan ketepatan pengkelasan kepada ketepatan yang boleh diterima kerana pengubahsuaian yang teruk pada sifat- sifat spektrum dan stastistik disebabkan oleh jerebu.

ACKNOWLEDGEMENTS

Alhamdulillah, I am thankful to almighty Allah Subhanahuwataala, the Most Gracious and the Most Merciful. His infinite mercy has guided me to complete this MSc. work. May peace and blessing of Allah be upon His Prophet Muhammad Sallallahualaihiwassalam. Who has been sent by Him as a mercy and blessing for the entire universe. I express my sincere gratitude to Dr. Asmala Bin Ahmad, my main supervisor for his never ending support and guidance toward the completion of this thesis. Thank you. My gratitude also goes to Associate Professor Dr. Burhanuddin Bin Mohd Aboobaider for all the motivation and help. Besides that, I am thankful to my colleagues at the Centre for Advance Computing Technology (C-ACT) and particularly Optimization, Modelling, Analysis, Simulation and Scheduling (OptiMASS) research group for the useful ideas and comments. I am most grateful to the Malaysian Remote Sensing Agency, Malaysian Meteorological Department and Department of Environment, Malaysia for providing the data.

TABLE OF CONTENTS

PAGE

DECLARATION		
APPROVAL		
DEDICATION		
ABSTRACT		**i**
ABSTRAK		**ii**
ACKNOWLEDGEMENTS		**iii**
TABLE OF CONTENTS		**iv**
LIST OF TABLES		**vii**
LIST OF FIGURES		**ix**
LIST OF APPENDIX		**xiii**
LIST OF ABBREVIATIONS		**xiv**
CHAPTER		
1.	**INTRODUCTION**	**1**
1.1	Research Background	1
1.2	Problem Statement	5
1.3	Research Objectives	6
1.4	Research Questions	6
1.5	Research Scope	8
1.6	Significance of Study	9
1.7	Thesis Plan	10
2.	**LITERATURE REVIEW**	**11**
2.1	Introduction	11
2.2	Remote Sensing	12
2.2.1	Active Satellite Sensor System	14
2.2.2	Passive Satellite Sensor System	15
2.3	Landsat Data	16
2.3.1	Landsat History	16
2.4	Land Cover Classification	21
2.4.1	Unsupervised Classification	22
2.4.2	Supervised Classification	22
2.5	Factors Affecting Classification	26
2.6	Haze	27
2.6.1	Visibility	27
2.6.2	Air pollution index	28
2.6.3	Path Radiance Concept	30
2.7	Previous Studies	31
2.7.1	Radiative Transfer Method	31
2.7.2	Image Based Method	33

2.8	Summary	38
3.	**RESEARCH METHODOLOGY**	**39**
3.1	Introduction	39
3.2	Data Preparation and Pre-processing	41
	3.2.1 Data Used and Image Processor	42
	3.2.2 Data Description and Study Location	43
	3.2.3 Data Subset and Calibration	46
	3.2.4 Data Registration	47
	3.2.5 Cloud, Cloud Shadow and Water Masking	47
3.3	Data Classification and Accuracy Assessment	50
	3.3.1 ROI Selection and Data Classification	50
	3.3.2 Accuracy Assessment of the Original Data	55
3.4	Haze Modelling and Simulation	60
	3.4.1 The Hazy Band	61
3.5	Classification and Accuracy Assessment of Hazy Datasets	62
3.6	Summary	66
4.	**DEVELOPMENT AND PERFORMANCE EVALUATION OF HAZE REMOVAL TECHNIQUE**	**67**
4.1	Introduction	67
4.2	Haze Removal Concept	67
4.3	Development of Haze Removal Technique	69
	4.3.1 PIF Identification	70
	4.3.2 Linear Regression Analysis and Haze Mean Subtraction	72
	4.3.3 Spatial Filtering	76
4.4	Accuracy Assessment	80
	4.4.1 Land Cover Classification	81
	4.4.2 Classification Accuracy	82
4.5	Summary	86
5.	**APPLICATION OF HAZE REMOVAL TECHNIQUE ON REAL HAZY DATA**	**87**
5.1	Introduction	87
5.2	Data Information	89
5.3	Methodology	93
	5.3.1 Data Pre-processing	94
	5.3.2 Haze Removal	94
5.4	Performance Evaluation	98
	5.4.1 Accuracy Assessment of Clear Data	98
	5.4.2 Accuracy Assessment of Hazy Data	103
	5.4.3 Analysis and Discussion	106
5.5	Summary	117

6.	**CONCLUSION AND RECOMMENDATIONS**	**118**
6.1	Conclusion	118
6.2	Suggestion on Future Works	120
REFERENCES		**121**
APPENDIX		**130**

LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Summary on correlation of problem statements, research objectives and research question and anticipated outcome	7
2.1	Wavelength and its description (GSFC, 2014)	13
2.2	Launch and decommissioned dates and type of sensor or Landsat series (USGS, 2012)	17
2.3	Band information for MMS and RBV (USGS, 2012)	19
2.4	Band information for TM and ETM+ (USGS, 2012)	19
2.5	Band information for OLI and TIRS (USGS, 2012)	20
2.6	Haze and visibility chart from the Malaysian Meteorological Department (MetMalaysia, 2015)	28
2.7	API range, status, level of pollution and health measures from Department of Environment (DOE, 2015)	29
3.1	Landsat TM spectral range and post calibration dynamic ranges	46
3.2	The Jeffries-Matusita distance (J-M distance) to measure the separability between classes	53
3.3	Confusion matrix for SVM classification (a) in pixels, (b) in percentage (%) and (c) producer accuracy	58
3.4	Overall accuracy and kappa coefficient for classified hazy dataset	66

4.1 The gain, offset and r^2 values obtained from regression analyses of the PIF values from the clear and hazy datasets (i.e. 18, 16, 14, 12, 10, 8, 6, 4, 2, and 0 km visibility) for bands 1, 2, 3, 4, 5 and 7 75

4.2 Band 1 and radiance scatter plot of a hazy data in (a) original form (b) after haze mean subtraction and (c) after haze mean subtraction and average filtering (3x3) for 18 km, 10 km and 2 km visibility 79

4.3 Best fit kernel size and filter type for respective visibility. The accuracy of hazy data before and after haze removal was presented side by side for comparison 85

5.1 Information of the hazy and clear data 90

5.2 Confusion matrix of the clear image in terms of (a) pixels, (b) percent and (c) producer accuracy for each class in terms of pixels and percentages 102

5.3 Confusion matrix of the classified image of hazy data with respect to the classified image of the clear data in terms of (a) pixels, (b) percent and (c) producer accuracy for each class 105

5.4 Confusion matrix of the classification after haze removal with respect to the classification of the clear data in terms of (a) pixels, (b) percent and (c) producer accuracy. 107

5.5 Minimum, maximum, mean and standard deviation for forest, oil palm and urban: (a) clear data, (b) hazy data and (c) after haze removal data 114

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	Precision farming concept (MRSA, 2011)	4
1.2	Chart for research scope	9
2.1	Writing organisation for Chapter 2	11
2.2	Range of electromagnetic spectrum (GSFC, 2014)	12
2.3	Active sensors emitting its own energy (GrindGIS, 2015)	14
2.4	Passive sensors detect reflected sunlight and thermal energy from Earth (GrindGIS, 2015)	16
2.5	Timeline of Landsat series (USGS, 2012)	17
2.6	Illustration of basic element for ANN (McCulloch and Pitts, 1943)	24
2.7	Basic idea of SVM (Vapnik, 1995)	25
2.8	Path radiance contribution to satellite signals during hazy conditions (Ahmad and Quegan, 2014)	31
3.1	Overall research workflow of the study	40
3.2	Flowchart to identify the effects of haze on remote sensing data properties	41
3.3	Location of the study area.	44
3.4	Reference map and Landsat 5 data	45
3.5	Landsat bands 3, 2 and 1 assigned to red, green and blue channel (left) and bands 4, 5 and 3 assigned to red, green and blue channel (right)	49

3.6 The mask band (left) and the Landsat data after masking process (right) 49
3.7 Landsat data bands 3, 2 and 1 assigned to red, green and blue channel with (a) training ROI and (b) reference pixels for oil palm after stratified random pixels 51
3.8 Pixels under original ROI polygons for oil palm (top) and closer view of Figure 3.7 with 50% and 25% of those pixels selected to be the training (middle) and reference pixels (bottom) 52
3.9 SVM classification of haze-free Landsat data dated 11 February 1999 55
3.10 Process of integrating haze layer to clear image 62
3.11 Hazy dataset (bands 3, 2 and 1 assigned to red, green and blue channel) on the left and classified image of the hazy dataset on the right for (a) 20 km (clear), (b) 10 km, (c) 6 km (d) 2 km and (e) 0 km visibility 64
3.12 Overall classification accuracy versus visibility 65
4.1 Flow chart for haze removal technique 70
4.2 Landsat bands 4, 5 and 3 assigned to red, green and blue channel of Klang, Selangor, Malaysia. (b), (c) and (d) are an enlarged version of PIF location in (a) from Google Maps 72
4.3 Scatterplots of PIF values from the 20 km dataset versus PIF values from the hazy datasets for visibilities 18, 10, 6, 2, and 0 km 74
4.4 An example of average filtering using a 3x3 kernel on anonymous pixels: (a) pixel values before average filtering and (b) pixel values after average filtering 77
4.5 An example of median filtering using a 3x3 kernel: (a) pixel values before median filtering and (b) pixel values after median filtering 78
4.6 Displays of classification for after removal data for 12 km visibility 81

4.7	Classification accuracy against visibility for average filtering	82
4.8	Classification accuracy against visibility for Gaussian filtering	83
4.9	: Classification accuracy against visibility for median filtering	83
4.10	Classification accuracy versus visibility	85
5.1	Flowchart of haze removal and accuracy assessment	88
5.2	Landsat 8 dated from 30th May 2015 and 19th September 2015 used as (a) clear and (b) hazy data respectively	92
5.3	Hazy data (a) and clear data (b) with bands 2,3 and 4 assigned to red, green and blue channel (left) and bands 5,6 and 4 assigned to red, green and blue channel (right)	93
5.4	Masks (left) and the corresponding haze segments (right) for (a) severe haze and (b) less severe haze	95
5.5	Display of bands 2, 3 and 4 assigned to red, green and blue channel for (a) clear data, (b) hazy data and (c) after haze removal data together with horizontal profile of the cross section	98
5.6	Image interpretation of Google Map image (GoogleMaps, 2015)	100
5.7	(a) Bands 4, 3 and 2 of the clear data assigned to red, green and blue channel (b) the classified image of the clear data	101
5.8	Classified image of (a) clear (b) hazy data (c) hazy data after haze removal	103
5.9	Subsetted area of classified image of (a) clear data (b) hazy data (c) after haze removal data	104
5.10	Mean radiances of Landsat bands 2, 3, 4, 5, 6 and 7 for forest, oil palm and urban: (a) clear data (b) hazy data and (c) hazy data after haze removal	109
5.11	Standard deviations of Landsat bands 2, 3, 4, 5, 6 and 7 for forest, oil palm and urban: (a) clear data (b) hazy data and (c) restored data	111

5.12 Maximum radiance value of each band for forest, oil palm and urban: (a) clear data (b) hazy data and (c) restored data 112
5.13 Minimum radiance value of Landsat bands 2, 3, 4, 5, 6 and 7 for forest, oil palm and urban: (a) clear data (b) hazy data and (c) restored data 113

LIST OF APPENDIX

APPENDIX	TITLE	PAGE
A	Data Justification	130

LIST OF ABBREVIATIONS

6SV1	Second Simulation of a Satellite Signal in The Solar Spectrum, Vector Version 1
ACORN	Atmospheric Correction Now
API	Air Pollution Index
ATCOR	Atmospheric Correction
CO	Carbon Monoxide
DEM	Digital Elevation Model
DN	Digital Number
DOE	Department of Environment, Malaysia
DOS	Dark Object Subtraction
ELM	Empirical Line Method
ENVI	Environment for Visualising Images
EROS	Earth Resources Observation System
ETM+	Enhanced Thematic Mapper Plus
FLAASH	Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes
GCP	Ground Control Point
GIS	Geographical Information System
GPS	Global Positioning System
GR	Ground Reference

HOT	Haze Optimized Transformation
IR	Infrared
JuPEM	Department of Survey and Mapping, Malaysia
L1T	Level 1 Terrain Corrected
MOSTI	Ministry of Science, Technology and Innovation
MRSA	Malaysian Remote Sensing Agency
MSS	Multi Spectral Scanner
NO_2	Nitrogen Dioxide
NASA	National Aeronautics and Space Administration
O_3	Ozone
OLI	Operational Land Imager
PCA	Principle Component Analysis
PIF	Pseudo Invariant Feature
PM_{10}	Particulate Matter Less than or Equal to 10 µm
RBV	Return Beam Vidicon
RMSE	Root Mean Square Error
ROI	Region of Interest
SAR	Synthetic Aperture Radar
SLC	Scan Line Corrector
SO_2	Sulphur Dioxide
SVM	Support Vector Machine
SWIR	Shortwave Infrared
TIRS	Thermal Infrared Sensors
TSS	Total Suspended Solid

USA	United State of America
USGS	United State Geological Survey
UTC	Universal Time Coordinator
UTM	Universal Transverse Mercator
UV	Ultra Violet
VI	Vegetation Index
WGS84	World Geodetic System 1984

CHAPTER 1

INTRODUCTION

1.1 Research Background

Approximately, only 30% of the earth surface is land while the other 70% is water. Nevertheless, land area has been a living place for almost 7 billion world population. Besides that, 99.7% of the world‘s food to human being comes from land while only 0.3% comes from oceans and other aquatic ecosystems (Pimentel and Wilson, 2004). With regards to this, there is a sense of urgency to monitor how human makes use land area to fulfil their necessities. For such purpose, there have been efforts to monitor land cover/land use over land using various tools. Initially land survey is carried out on foot where land cover/land use information is written manually on papers. By using this approach, a huge amount of information can be obtained however, it consumes a lot of time, man power, besides logistically expensive particularly for large areas.

With technology advancement, in mid-1800s, camera was invented where picture was taken by photographers not only from ground but also from balloons in attempt to get a larger coverage of land. The first unmanned camera was later introduced in the end of 1800s where camera was mounted on a kite and later in the early 1900s on an aircraft in which was known as aerial photography. Aerial photography has been used to monitor land cover/land use for more than 50 years. Although aerial photography has successfully facilitated the process of obtaining land information nevertheless, the maintenance cost for the sensor and the airplane are very expensive especially for continuous implementation. Satellite remote sensing was only been used as an alternative to aerial photograph in 1972 when the first land satellite, known as ‗Landsat 1‗ was launched by NASA (National Aeronautics and Space Administration), USA with the primary mission to map global land cover. Since then, remote sensing data have been one of the most important tools not only for monitoring but also mapping land uses/covers. Land use and land cover map is important for various decisions making and planning. The information is widely used by users ranging from students, researchers, engineers technical workers to policy makers.

Among the most crucial use of satellite remote sensing at regional and national level is for monitoring land cover/use change, deforestation, natural disaster and gazette the land boundary. This is primarily important for many countries including Malaysia that is fast developing and experiencing vast changes in land development and urban modification. If the land use is not monitored and managed properly, it may result in devastating disaster such as landslide, flash flood in urban area, land degradation, destruction tropical land forest and loss of biodiversity.

Before further discussion, we first need to clarify the definition of land use and land cover. Land use is referred as how human exploit the land properties. It includes land use for modification of management of land for agricultural, urbanization, forestry and forest conservation. While land cover describes physical material on the earth, it can be natural or planted vegetation, urban infrastructure, water, or anything that can be identified on the earth surface (Mohsin, 2014). In other words, land use indicates how people utilize land, while land cover indicates physical land type, therefore unlike land cover, land use cannot be determined from satellite remote sensing. Satellite remote sensing can determine land cover which in turn can infer land use that present at a particular area.

Land cover maps produced by remote sensing imagery have been widely used to fulfil global, regional and national needs. Malaysia is also not left behind in producing her own national land cover maps. It has been produced by the Department of Agriculture

(DOA) since 1966 (Mahmood et al., 1997) and later through a collaboration work between DOA and Malaysian Remote Sensing Agency (MRSA) or formerly known as Malaysian Centre for Remote Sensing (MACRES), under the government's Ministry of Science, Technology and Innovation (MOSTI).

The availability of land cover map has trigged research projects nationally and regionally particularly with regards to resources and environmental management. One of the major national projects coordinated by MRSA is precision farming. Precision farming is aimed to increase crop production by systematically monitoring plant growth, yield condition, soil moisture, water irrigation, and weather condition. This is done by integrating remote sensing, GPS and GIS technologies with ground farming facilities. In precision farming, plant phenology data can be systematically collected, modelled and eventually crop yield for every harvesting season can be predicted. Also, plant necessities such as fertilizers, pesticides and water can be optimally consumed to safeguard plant vigour. The precision farming concept is illustrated in Figure 1.1. In Malaysia, precision farming has been implemented to paddy field and oil palm. In paddy farming, there is a need for continuous satellite data supply in short period of time as cultivation period of paddy is between 110 to 135 days from sowing seeds until harvesting. However, in Malaysia data acquisition using remote sensing devices tends to be interrupted by environmental factors, particularly haze. When haze is too severe, satellite data for certain period of time cannot be obtained, therefore causing problems in data analysis and interpretation tasks.