Document image analysis of Balinese palm leaf manuscripts

Made Windu Antara Kesiman

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE EUCLIDE

Laboratoire Informatique, Image et Interaction (L3i)

THÈSE présentée par :

Made Windu Antara KESIMAN

soutenue le : **5 Juillet 2018**

pour obtenir le grade de : **Docteur de l'université de La Rochelle**

Discipline : **Informatique et Applications**

Document Image Analysis of Balinese Palm Leaf Manuscripts

JURY :

Andreas FISCHER	Professeur, University of Applied Sciences and Arts Northwestern Switzerland (Suisse), Examinateur
Elisa Barney SMITH	Professeur, Boise State University (États-Unis), Rapporteur
Josep LLADÓS	Professeur, Universitat Autònoma de Barcelona (Espagne), Rapporteur
Gede Rasben DANTES	Maître de conférences, Universitas Pendidikan Ganesha (Indonesie), Examinateur
Jean-Christophe BURIE	Professeur, Université de La Rochelle (France), Directeur de thèse
Jean-Marc OGIER	Professeur, Université de La Rochelle (France), Co-Directeur de thèse, Président du jury
Philippe GRANGÉ	Maître de conférences, Université de La Rochelle (France), Co-Directeur de thèse

Matur suksma Ida Sang Hyang Widhi Waça ...
To my parents *Bapak dan Ibu*, my sister *Mbok Wik*, my wife *Dek Susi*, my daughter *Utu Kalpika* and my son *Dek Genta* ...

Acknowledgements

This dissertation is actually a fruitful collaborative work from many peoples who were willing to help and to share their knowledge and times during almost four years of my research work. And in this special page, I would like to thank them all.

First of all, I would like to thank Prof. Elisa Barney Smith and Prof. Josep Lladós to have reviewed this dissertation and for sharing constructive comments and very rich discussions. I thank also Prof. Andreas Fischer and Dr. Gede Rasben Dantes for accepting to be part of the jury.

My sincere thanks to my supervisors Prof. Jean-Christophe Burie, Prof. Jean-Marc Ogier and Dr. Philippe Grangé for their precious expertise, friendship, dynamism, understanding, time and energy to supervise this work at the Laboratoire Informatique, Image et Interaction (L3i) of the Université de La Rochelle (ULR). I thank also Prof. Mohamed Yacine Ghamri Doudane as Director of the L3i for accepting me at the lab with all supports during my stay and all facilities to run this work in an open vision and environment.

I would like to thank Museum Gedong Kertya, Museum Bali, I Made Muliarta, Dona Valy, I Gusti Ngurah Made Agus Wibawantara, I Made Gede Sunarya, Hendro, Credo, and all families in Bali, Indonesia, for providing the samples of palm leaf manuscripts. I also thank the students from the Laboratory of Cultural Informatics (LCI), the Department of Informatics Education and the Department of Balinese Literature, University of Pendidikan Ganesha, for the helps in the ground truthing process for this research project. This work is supported by the DIKTI BPPLN Indonesian Scholarship Program and the STIC Asia Program implemented by the French Ministry of Foreign Affairs and International Development (MAEDI).

My stay at the lab would not be pleasant without Nam, Elodie, Bao, Christophe, Hai, Bouziane, Florien, Mohammad and all young researchers who stay in and come to room 121bis. I thank you all for friendship, togetherness, time, stories, humor and laughs that we have shared. Many helps from Kathy, Erlandri, Geneviève, Muzzamil and Dom, Isabelle and Jennifer from Ecole Doctorale, are also priceless that I would like to thank them all.

Last but not least, I would like to say *matur suksma Ida Sang Hyang Widhi Waça*, and this dissertation will not be finished without the love from my family, Bapak dan Ibu, Mbok Wik, Om Semut, Adhira, Auryn, Adhyasta, Mek Ade, Bapak dan Memek Mandor, and my wife Dek Susi with my daughter Utu Kalpika and my son Dek Genta ...

Abstract

The collection of palm leaf manuscripts is an important part of Southeast Asian people's culture and life. Following the increasing of the digitization projects of heritage documents around the world, the collection of palm leaf manuscripts in Southeast Asia finally attracted the attention of researchers in document image analysis (DIA). The research work conducted for this dissertation focused on the heritage documents of the collection of palm leaf manuscripts from Indonesia, especially the palm leaf manuscripts from Bali. This dissertation took part in exploring DIA researches for palm leaf manuscripts collection. This collection offers new challenges for DIA researches because it uses palm leaf as writing media and also with a language and script that have never been analyzed before. Motivated by the contextual situations and real conditions of the palm leaf manuscript collections in Bali, this research tried to bring added value to digitized palm leaf manuscripts by developing tools to analyze, to transliterate and to index the content of palm leaf manuscripts. These systems aim at making palm leaf manuscripts more accessible, readable and understandable to a wider audience and, to scholars and students all over the world. This research developed a DIA system for document images of palm leaf manuscripts, that includes several image processing tasks, beginning with digitization of the document, ground truth construction, binarization, text line and glyph segmentation, ending with glyph and word recognition, transliteration and document indexing and retrieval. In this research, we created the first corpus and dataset of the Balinese palm leaf manuscripts for the DIA research community. We also developed the glyph recognition system and the automatic transliteration system for the Balinese palm leaf manuscripts. This dissertation proposed a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts. The proposed scheme consists of six tasks: the text line and glyph segmentation, the glyph ordering process, the detection of the spatial position for glyph category, the global and categorized glyph recognition, the option selection for glyph recognition and the transliteration with phonological rules-based machine. An implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscript is proposed. The adaptation of a segmentation-free LSTM-based transliteration system with the generated synthetic dataset and the training schemes at two different levels (word level and text line level) is also proposed.

Keywords: palm leaf manuscripts, Balinese script, document image analysis, dataset, binarization, text line segmentation, glyph recognition, transliteration

Résumé

Les collections de manuscrits sur feuilles de palmier sont devenues une partie intégrante de la culture et de la vie des peuples de l'Asie du Sud-Est. Avec l'augmentation des projets de numérisation des documents patrimoniaux à travers le monde, les collections de manuscrits sur feuilles de palmier ont finalement attiré l'attention des chercheurs en analyse d'images de documents (AID). Les travaux de recherche menés dans le cadre de cette thèse ont porté sur les manuscrits d'Indonésie, et en particulier sur les manuscrits de Bali. Nos travaux visent à proposer des méthodes d'analyse pour les manuscrits sur feuilles de palmier. En effet, ces collections offrent de nouveaux défis car elles utilisent, d'une part, un support spécifique : les feuilles de palmier, et d'autre part, un langage et un script qui n'ont jamais été analysés auparavant. Prenant en compte, le contexte et les conditions de stockage des collections de manuscrits sur feuilles de palmier à Bali, nos travaux ont pour objectif d'apporter une valeur ajoutée aux manuscrits numérisés en développant des outils pour analyser, translittérer et indexer le contenu des manuscrits sur feuilles de palmier. Ces systèmes rendront ces manuscrits plus accessibles, lisibles et compréhensibles à un public plus large ainsi que pour les chercheurs et les étudiants du monde entier. Cette thèse a permis de développer un système d'AID pour les images de documents sur feuilles de palmier, comprenant plusieurs tâches de traitement d'images: numérisation du document, construction de la vérité terrain, binarisation, segmentation des lignes de texte et des glyphes, la reconnaissance des glyphes et des mots, translittération et l'indexation de document. Nous avons ainsi créé le premier corpus et jeu de données de manuscrits balinais sur feuilles de palmier. Ce corpus est actuellement disponible pour les chercheurs en AID. Nous avons également développé un système de reconnaissance des glyphes et un système de translittération automatique des manuscrits balinais. Cette thèse propose un schéma complet de reconnaissance de glyphes spatialement catégorisé pour la translittération des manuscrits balinais sur feuilles de palmier. Le schéma proposé comprend six tâches: la segmentation de lignes de texte et de glyphes, un processus de classification de glyphes, la détection de la position spatiale pour la catégorisation des glyphes, une reconnaissance globale et catégorisée des glyphes, la sélection des glyphes et la translittération basée sur des règles phonologiques. La translittération automatique de l'écriture balinaise nécessite de mettre en œuvre des mécanismes de représentation des connaissances et des règles phonologiques. Nous proposons un système de translittération sans segmentation basée sur la méthode LSTM. Celui-ci a été testé sur des données réelles et synthétiques. Il comprend un schéma d'apprentissage à deux niveaux pouvant s'applique au niveau du mot et au niveau de la ligne de texte.

Mots-clés: manuscrits sur feuilles de palmier, écriture balinaise, analyse d'images de documents, jeu de données, binarisation, segmentation de ligne de texte, reconnaissance de glyphe, translittération

Table of Contents

List of Figures

List of Tables

List of Algorithms

Glossary of Terms

lontar palm leaf manuscript in Balinese language. 15, 16, 17, 28, 33

pengerupak small knife to scribe the palm leaf. 28, 39

List of Acronyms

ALCM Adaptive Local Connectivity Map. viii, xiii, 61, 67, 68, 158, 162

AMADI Ancient Manuscripts Digitization and Indexation. ix, xi, xii, xix, 3, 5, 6, 19, 24, 91, 107, 108, 109, 119, 125, 129, 145, 166, 181, 192, 194, 198, 207, 225, 226, 225, 227, 225, 228, 225, 229, 225, 230, 225

APP Adaptive Partial Projection. viii, xiii, xvi, 61, 62, 158, 160, 158

BLSTM Bidirectional Long Short-Term Memory. 88, 179

CER Character Error Rate. 171, 184

CNN Convolutional Neural Network. ix, xvii, 49, 58, 81, 82, 83, 84, 166, 168, 169

DIA Document Image Analysis. vii, viii, 4, 5, 6, 8, 9, 13, 33, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 46, 49, 50, 51, 52, 51, 53, 54, 53, 54, 55, 56, 55, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 85, 87, 88, 89, 88, 90, 113, 129, 139, 155, 190, 192

DIBCO Document Image Binarization Competition. xiii, 57, 60, 96, 141, 145, 153

DR Detection Rate. xviii, 158, 162

FM F-Measure. xviii, 141, 153, 154

FN False Negative. 141

FP False Positive. 141

GIC Génie Informatique et Communication. 3

HoG Histogram of Gradient. viii, xiv, 81, 124, 164, 166, 169

ICDAR International Conference on Document Analysis and Recognition. 198

ICFHR International Conference on Frontiers in Handwriting Recognition. viii, ix, xvi, 6, 49, 58, 60, 84, 110, 153, 155, 164, 168, 192

IHCR Isolated Handwritten Character Recognition. 75, 88, 124, 166, 169

ITC Institute of Technology of Cambodia. 3

k-NN k-Nearest Neighbor. 124, 164, 168, 169

L3i Laboratoire Informatique Image Interaction. 3

LCI Laboratory of Cultural Informatics. 3

LSTM Long Short-Term Memory. ix, 7, 50, 58, 85, 87, 88, 176, 179, 181, 183, 190, 192

MAEDI Ministère des Affaires Etrangères et du Développement International. 3

MSE Mean Square Error. 141

NN Neural Network. 169

NPW Neighborhood Pixels Weights. xiv, xv, 75, 81, 80, 124, 125, 124, 164, 166, 169

NRM Negative Rate Metric. 141, 149

NRR Number Recognized Result. 170, 171

NSG Number Segments Ground Truth. 170, 171

NSO Number Segments Overlapped. 170, 171

NSR Number Segments Result. 170, 171

OCR Optical Character Recognition. xv, 13, 43, 54, 56, 61, 75, 76, 85, 88, 96, 130, 134, 135, 136, 137, 174, 181, 183

PPR Precision Pattern Rate. 171

PR Pattern Rate. 171

PSNR Peak Signal to Noise Ratio. xvi, 141, 149, 151

RAID Robotics, Artificial Intelligence, and Digital image. 3

RNN Recurrent Neural Network. 85

RPR Recall Pattern Rate. 171

SR Segmentation Rate. 170, 171

SRR Segmented Recognition Rate. 170, 171

STIC Sciences et Technologies de l'Information et de la Communication. 3

TN True Negative. 141

TP True Positive. 141

UFL Unsupervised Feature Learning. 83, 169

ULR Université de La Rochelle. 3

UNDIKSHA Universitas Pendidikan Ganesha. 3

UNPAD Universitas Padjadjaran. 3

Chapter 1

Introduction

This chapter brings the reader into the scopes of this dissertation by introducing the research context and understanding the challenges, objectives and motivations of the research. The research hypothesis is also clearly stated in this chapter. This chapter ends with a summary of contributions of this dissertation.

1.1 Research Contexts

Every nation and country in the world certainly has its own cultural and historical timeline which of course also becomes part of the world history and civilization as a whole. The history and the world civilization are often stored, or more precisely are recorded in the form of written documents. The beginning of recording activities into this form of written document in the history of human civilization is known as a phase of transition from the prehistoric period to the historical period. Generally described, prehistory is the period of human activity before the invention of writing systems, a time period where no written records exist or period before writing systems were introduced[1]. And since the beginning of the historical period, people tend to express and record all important aspects of their lives, from spiritual to economic topics, in a written document. Everyday life's features were less often recorded. The term heritage document refers to all the relics of these written documents created by hand or by machine, made from the earliest period until the most recent one. All writing cultures from the ancient times until the modern times amassed huge collections of heritage documents [20].

The main context and purpose of the preservation project for heritage documents is actually to preserve and save all the valuable contents of the document. However, preservation measures are more often related only for saving the physical condition of the material of the document. But these two things are of course closely related. It is just worth noting that the preservation steps of ancient documents are not merely about physical or chemical treatment for the documents, such as cleaning documents, or designing proper storage for documents. The efforts of transferring the contents of the document into other media are also actually important preservation steps to be done immediately. And since the world entered the era of digital media in the beginning of 21st century, the efforts to digitize important documents are increasingly campaigned around the world. Digital media technology acts like a very strong spike rock supporting efforts to save and preserve heritage documents. And following the increasing development of the preservation project for heritage documents around the world, the collection of palm leaf manuscripts in Southeast Asia finally attracted the attention of researchers in Document Image Analysis (DIA). Digitization and indexing projects for palm leaf manuscripts were proposed. For example, the digitization project for palm leaf manuscripts from Cambodia[2] and Thailand [21, 22].

Within the scope of the heritage documents, the research work conducted for this dissertation will mainly focus on the heritage documents of the collection of palm leaf manuscripts from Indonesia, especially the palm leaf manuscripts from Bali. In some parts, the collection of palm leaf manuscripts from Sunda (West Java - Indonesia) and the Khmer palm leaf manuscripts from Cambodia will be also described.

The collection of palm leaf manuscripts became an integral part of Southeast Asian people's culture and life. For example, the existence of palm leaf manuscript collections in Bali, Indonesia has also been known from long time ago. But the discovery of palm leaf manuscript was generally not seen as great inventions like the papyrus documents and tablets from the Egyptians. This is because most of the collections have been easily found and collected in Balinese temples that are still standing and functioning to date and also

[1] http://pediaa.com/difference-between-history-and-prehistory/
[2] http://www.khmermanuscripts.org/

in many Balinese households. But over the time, the collection of such important heritage documents seemed to be forgotten. It is as if no one is interested and needs those kinds of manuscripts. For decades, most of the collection of such important documents stays in silence, it is never touched and is never opened again.

But nowadays, the interest in the content of Balinese palm leaf manuscripts from the Balinese peoples in particular and also from many researchers around the world is actually appearing again. The Balinese peoples now really have a great interest to know and to understand what has been written in their palm leaf manuscripts. The researchers from many scientific research backgrounds are coming to Bali with the expected goals to learn and to find their research domain resources in the collection of the Balinese palm leaf manuscripts. But, they are mostly facing the same challenging situations and real conditions when the access to the collections and to the content of palm leaf manuscripts are limited for some reasons as follows:

1. Physical condition and the fragility of the palm leaf manuscripts. Many degraded old palm leaf manuscript collections are not allowed to be touched and to be opened again. Even though some of them are already rewritten into new palm leaf materials, but most of them are still not indexed and listed in a complete catalog.

2. Limited access to the private family collection. Although some museums or cultural agencies or institutions collect and provide most of the principal palm leaf manuscript collections, a very large number of collections are still kept by private families. And it is not easy to get an access or a permission to these private collections. The religious, sacred and spiritual reasons have to be fully considered by everyone who wants to access to the collections.

3. Linguistic difficulties. Although the existence of ancient palm leaf manuscripts in Southeast Asia is very important, unfortunately, the access to their content, especially in Bali, is limited due to linguistic difficulties. Even for most of the Balinese peoples, they never read and are not able to read any palm leaf manuscript. This is because in Bali, peoples still speak in Balinese, but they are now writing in Latin script. Balinese script is no longer used in the writing activities. It is still used in some religious events, but only by a small number of Balinese philologists. The young generation tends to forget and does not learn how to write the Balinese script. And because of this main obstacle, it is almost impossible for them to be able to understand the valuable content of the palm leaf manuscripts.

4. Difficulties in searching for a certain collection in the catalog of museums or cultural agencies or institutions. Finding a certain content of the manuscript from the thousand collections by using some keywords are still not straightforward and it will be very time consuming to be done from the manual catalog. Moreover, the collections of transliterated texts of the manuscripts are not complete. Some digitized version of the manuscripts are actually already available. But most of them are done only for digital data backup of the manuscripts, they are not accompanied by their associated transliterated text, or the complete information about the content of the manuscripts.

Those challenging situations lead multidisciplinary scientific challenges to the context of this research with socio-cultural, philological and linguistic aspects and the need of document image analysis. (Figure 1.1). This dissertation is conducted under the scheme

Figure 1.1: Research Contexts

of the AMADI (Ancient MAnuscripts Digitization and Indexation) Project[3], in the contexts and scopes of efforts to save, preserve, and disseminate important contents and knowledge stored in Balinese palm leaf manuscript collections. The AMADI Project is an international collaborative project between three countries, which are France as the leader of the project, Indonesia and Cambodia as two Asian country partners. This work is supported by the STIC Asia Program[4] implemented by the French Ministry of Foreign Affairs and International Development (MAEDI). This project involves four research institutions, which are Laboratory L3i (Laboratoire Informatique Image Interaction)[5] of Université de La Rochelle (ULR)[6], France, Laboratory LCI (Laboratory of Cultural Informatics)[7] of Universitas Pendidikan Ganesha (UNDIKSHA)[8], Bali, Indonesia, Laboratory RAID (Robotics, Artificial Intelligence, and Digital image) and Center for Sundanese Culture Studies of Universitas Padjadjaran (UNPAD)[9], Bandung, Indonesia, and Laboratory GIC (Génie Informatique et Communication) of Institute of Technology of Cambodia (ITC)[10], Phnom Penh, Cambodia. This cross-disciplinary project team brings together researchers in computer science, philology of ancient literary text, and social human sciences.

1.2 Motivations and Objectives

Motivated by the contextual situations and real conditions of the palm leaf manuscript collections in Bali, this research tries to bring added value to digitized palm leaf manuscripts by developing tools to analyze, to transliterate and to index the content of palm leaf manuscripts. These systems will make palm leaf manuscripts more accessible, readable and understandable to a wider audience and to scholars and students all over the

[3]http://amadi.univ-lr.fr
[4]http://www.campusfrance/fr/Stic-asie
[5]http://l3i.univ-larochelle.fr
[6]http://www.univ-larochelle.fr
[7]http://pti.undiksha.ac.id/lci
[8]http://www.undiksha.ac.id
[9]http://www.unpad.ac.id
[10]http://www.itc.edu.kh

world. This research works not only to digitize the palm leaf manuscripts (Figure 1.2), but also to develop an automatic analysis, transliteration and indexing system for the manuscripts.

Technically, this research aims to develop a DIA system for document images of palm leaf manuscripts, that includes several image processing tasks, beginning with digitization of the document, ground truth construction, binarization, text line and glyph segmentation, ending with glyph and word recognition, transliteration and document indexing and retrieval. In line with motivations of the works, the objectives of this research are as follows.

1. To constitute the first corpus and dataset of the Balinese palm leaf manuscripts for the DIA research community. The new characteristics of palm leaf manuscripts provide a suitable challenge for testing and evaluation of robustness for the existing methods of document image analysis.

2. To develop the glyph recognition system and the automatic transliteration system for the Balinese palm leaf manuscripts. The glyph recognition system and also an automatic transliteration system are urgently needed for this kind of ancient manuscript collection. Using a glyph recognition system will help to transliterate these ancient documents and facilitate the next step to translate them to the current language, to give an access to the important information and knowledge in palm leaf manuscript. A transliteration engine for transliterating the Balinese script of palm leaf manuscript to the Latin-based script is one of the most demanding systems which has to be developed for the collection of palm leaf manuscripts.

3. To develop the automatic indexing and search engine system for the manuscript collections. With the help of the transliteration system, the automatic indexing and search engine system for the manuscript collections can be built. It will help the philologists to index and to access the content of the manuscripts quickly and efficiently. The syllables, words or any lexical-statistic can be generated and be indexed by the philologists, and it can be easily searched by everyone who needs some certain information from the manuscript's contents.

1.3 Research Hypothesis

The main hypothesis of this research work is based on the fact that due to high variety of the input material with the special characteristics and challenges possessed by the Balinese palm leaf manuscript collections, it will require a thorough adaptation of the DIA systems. Although there are some components of DIA system that are allegedly already generic to be applied to many types of documents, they will still need some specific adjustments to be able to be applied to these new types of documents. The solution of a DIA system as a taxonomy of the document processing steps which will be applied from the raw form of document images to the structured computer data is not unique and is not universal for all types of problems from different document collections. However, among DIA system's non-unique solutions, one specific solution can still be designed to deliver the most optimal DIA system's performance while still considering and taking into account the most suitable real condition found in that problem.

Figure 1.2: Four different palm leaf manuscript images for the AMADI Project

1.4 Contributions of this Dissertation

The main contributions of this dissertation are summarized below.

1. The first contribution is the creation of the Balinese palm leaf manuscript corpus, the design of the ground truth tools and protocols, and finally the presentation of the first Balinese palm leaf manuscript dataset for the DIA research as follow:

   - Collecting and digitizing the Balinese palm leaf manuscripts corpus (see Section 5.1 and 5.2).
   - Analyzing and proposing a specific scheme for the construction of the ground truth of binarized images [1] (see Sub Section 5.3.2).
   - Analyzing the subjectivity of the human intervention during the construction of the ground truth of binarized images and measuring quantitatively the ground truth variability of the binarized images [18] (see Sub Section 7.1.2.2 and 7.1.2.3).
   - Designing the overall scheme of ground truth construction and annotation protocols and presenting the AMADI_LontarSet, the first handwritten Balinese palm leaf manuscript dataset [15] (see Section 5.3).
   - Organizing competitions on the document image analysis tasks for Balinese Palm Leaf Manuscripts and Southeast Asian Palm Leaf Manuscripts for a wider DIA research communities, in the 15th and the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016 and 2018) [23]. After the competition, the dataset have been made publicly available for scientific use[11,12].

2. The second contribution consists of a number of experimental evaluations and empirical benchmarkings of commonly used DIA methods and algorithms for the palm leaf manuscript dataset as follows:

[11] http://amadi.univ-lr.fr/ICFHR2016_Contest/
[12] http://amadi.univ-lr.fr/ICFHR2018_Contest/

- Binarization: Experimenting and comparing several alternative well-known binarization algorithms, and in order to overcome the binarization problem on degraded and low quality palm leaf manuscript images, proposing a 'semi-local' concept to apply a powerful global binarization method on only precise local character area [1, 24] (see the methods in Sub Section 4.3.1 and 5.3.2.1, and the evaluations in Section 7.1).
- Text line segmentation: Investigating the performances of text line segmentation methods by conducting comparative experimental studies on the collection of Southeast Asian palm leaf manuscript images [11, 25, 24] (see the methods in Sub Section 4.3.2 and the evaluations in Section 7.2).
- Isolated glyph recognition: Investigating and evaluating some most commonly used features for character recognition, proposing and evaluating the combination of features, and implementing the supporting glyph recognition for the transliteration of Balinese script [14, 24, 16] (see the methods in Sub Section 4.3.3 and the evaluations in Section 7.3).
- Word transliteration: Evaluating the segmentation free LSTM based method for word transliteration of Southeast Asian palm leaf manuscript images [24] (see the methods in Sub Section 4.3.4, and the evaluations in Section 7.5).

3. The third contribution consists of developing a segmentation-based glyph recognition for the transliteration scheme of Balinese palm leaf manuscripts as follows:

   - Proposing a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts [16]. The proposed scheme consists of six tasks: the text line and glyph segmentation, the glyph ordering process, the detection of the spatial position for glyph category, the global and categorized glyph recognition, the option selection for glyph recognition and the transliteration with phonological rules-based machine. Detailed description of each task in this scheme is given in Chapter 6.
   - Proposing an implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscripts. A rule-based engine for performing transliterations is proposed [26]. The phonological rules are built and are formally defined based on the glyph recognition output. A rule-based engine for performing transliterations is proposed. This model consists of phonetic rules which are based on traditional linguistic study of Balinese transliteration. Detailed description of this proposition is given in Section 6.6.

4. The fourth contribution is the adaptation of a segmentation-free LSTM-based transliteration system of Balinese palm leaf manuscripts as follows:

   - Developing the automatic synthetic handwritten Balinese script generator. This application generates automatically and synthetically an image of Balinese script from a Latin text to simulate the degraded handwriting sample on a Balinese palm leaf manuscript.
   - Proposing and evaluating some adapted segmentation free training schemes for the transliteration of the Balinese script into the Latin script from palm leaf manuscript images. The generated synthetic dataset and the training schemes at two different levels (word level and text line level) are proposed. Detailed description of this scheme is given in Sub Section 7.5.3.2.1.

These contributions from this dissertation have led to some communications and publications listed in Appendix A.

1.5 Organization of this Dissertation

This dissertation is organized in eight chapters as follows:

- **Chapter 1** brings the reader into the scope of this dissertation by introducing the research context, objectives and motivations. The research hypothesis is also clearly stated in this chapter. This chapter is ended with a summary of contributions of this dissertation.
- **Chapter 2** presents a brief overview of the existence of heritage documents in general from the beginning of history of writing materials. It also describes the interests of heritage documents from some points of view. This chapter is also dedicated to the general presentation of the socio-cultural aspects of palm leaf manuscripts from Southeast Asia. This chapter exposes more specifically the unique characteristics of Balinese palm leaf manuscripts, the collections, the productions and the writing tools. The basic concept of Balinese language with syllabic script is presented in this chapter.
- **Chapter 3** describes and discusses the challenges for the development of a DIA system of Balinese palm leaf manuscripts in three parts. The first challenge is the socio-cultural aspects. The second challenge is the physical characteristic of the manuscripts, and the third challenge is the complexity of Balinese script.
- **Chapter 4** concentrates on the presentation of all existing methods for each task from the state-of-the-art of DIA system. The global overview of a DIA system and the ground truth construction were firstly given. The more detailed description of existing methods for each task in DIA's pipeline is then presented.
- **Chapter 5** presents the corpus and ground truth dataset of Balinese palm leaf manuscripts which are collected, constructed and used for all research works in this dissertation. This chapter describes the protocol design and the complete process from the manuscript digitization process until the dataset annotation process. The additional corpus and dataset from Khmer and Sundanese palm leaf manuscripts are also presented in this chapter.
- **Chapter 6** presents the proposed complete scheme of spatially categorized glyph Recognition for the transliteration of Balinese palm leaf manuscripts. This chapter presents the detailed description of each of the six tasks of the proposed scheme, the design of a segmentation technique customized for this respective problem of manuscript analysis and the flow of recognition and the option selection rules for glyph recognition. This chapter also describes the knowledge representation and phonological rules which are built for the transliteration engine of Balinese script.
- **Chapter 7** summarizes all experimental evaluations which have been done in this research work and discusses the results for each task and step in the DIA system for Balinese palm leaf manuscripts.
- **Chapter 8** finally gives some conclusions of the work presented in this dissertation by describing some limitations of the proposed system and the possible improvements for future work.

Chapter 2

Heritage Documents and Palm Leaf Manuscripts

This chapter presents a brief overview of the existence of heritage documents in general from the beginning of history of writing materials. It also describes the interests of heritage documents from some points of view. This chapter is also dedicated to the general presentation of the socio-cultural aspects of palm leaf manuscripts from Southeast Asia. This chapter exposes more specifically the unique characteristics of Balinese palm leaf manuscripts, the collections, the productions and the writing tools. The basic concept of Balinese language with syllabic script is also presented in this chapter.

2.1 General Description on the Existence of Heritage Documents

The invention of writing certainly marks the beginning of the existence of heritage documents around the world. Although initially the main focus of the invention of writing is about the material and the set of symbols or the scripts that humans can use to write, the next interests of writing is about the content that should and will be written. The combination of the existing materials, scripts and the content variations of writing generally provide and offer the points of interest of heritage documents as a critical and crucial object in document image analysis research.

2.1.1 A Brief History of Writing Materials

The invention of writing was triggered by the human desire to be able to convey the message to others in the distant space and different time, as well as to fulfill human tendency to store and preserve important information that may be forgotten. The evolution of the recording process in the form of a written document itself is a very important historical and cultural journey for world civilization from the very first time. It was influenced by three major historical axes, those are the invention of writing materials and tools, the use of the writing systems (symbols, alphabets, or scripts), and the evolution of spoken language to be written. The development of the script or alphabet and language eventually became an integral part of each other in the next historical period.

In general, regarding the historical timeline of the use of writing systems, it was found that in the beginning of the pictographic writing system, symbols were only used to represent objects. Then, the Egyptians introduced a writing system which is popularly known as hieroglyphs. It was also believed that long before the Greeks and Romans, Sumerian writing system and Egyptian hieroglyphs contain not only ideograms or logographics, but also phonetic signs. A simplistic taxonomy of types of writing systems can be categorized in five classes, such as Alphabetic, Syllabic, Logographic, Pictographic and Ideographic [20]. The writing system for the Sanskrit language and its derivatives follows the phonological alpha syllabic system. The principal implication of this types of writing systems for document recognition, more particularly as challenges for character or glyph recognition and text transliteration will be discussed in Sub Section 3.3.1 and 3.3.2.

Regarding the use of writing materials and tools, history records the discovery of important documents written on stone plates, clay plates or tablets, bark, skin, animal bones, ivory, tortoiseshell, papyrus, parchment which is a form of leather and is made of processed sheepskin or calfskin[1], copper and bronze metal plates, bamboos, palm leaves, and other materials for paper [20]. It is known that prehistoric men have begun drawing or painting or writing symbols on cave walls with natural dye materials and pigments they find, even using their own blood. Later, it was discovered that the Sumerians in the southern region of Mesopotamia began writing by printing symbols on wet clay plates or clay tablets[??]. This writing system is called cuneiform and this is considered as the first writing developed in world civilization history[??2]. Papyrus was produced as early as 3000 BCE in Egypt, and in ancient Greece and Rome. The word paper comes from this word of papyrus plant[1]. Even that clay tablets are almost indestructible because

[1]http://www.casepaper.com/company/paper-history
[2]http://www.historyworld.net/wrldhis

it was dried hard in the sun[2], papyrus offer many new advantages as a writing material. Papyrus has a flexible smooth surface and it accepts and retains ink without blur or smudge[2]. Papyrus were used longer than any other material in the history of written documents. During the ancient times, there was a huge production of clay tablets in Mesopotamia and papyrus in Egypt [20].

The choice of natural materials that can be used as a medium for document writing is strongly influenced by the geographical condition and location of a nation. For example, because geographically bamboo and palm trees are easily found in Asia, both types of materials are the first choice as writing material in the Asian continent. The thin strips of bamboo were used in China, and the palm leaf manuscripts were widely found in India and Southeast Asian countries[2]. In the Southeast Asian region, palm trees are ubiquitous and easy to find, therefore the Southeast Asian people choose to use palm leaves for writing media. In South and Southeast Asia, palm leaves were the most used medium until modern times [20].

Finally, paper was invented 2000 years ago in ancient China[3]. It is not quite certain whether the Chinese introduced paper to Indonesia. But the Europeans have brought paper since the beginning of the colonial period in Indonesia. And with the mass production of paper, the writing activity of the entire population of the world has increased sharply ever since. Major works in the various fields of science, art, culture and literature, and the latest discoveries of knowledge were written quickly. Therefore, the contents of these heritage documents are very important as evidence of world civilization history. And over time, the entire collection of documents written since the first time that the history of the world began to be recorded has now reached an extraordinary amount. These very valuable heritage documents are spread all over the world, and in many cases, these documents had been copied on the paper many times.

2.1.2 The Interests of Heritage Documents

There are some points of interest in working with the collections of heritage documents. Firstly, the huge quantity of documents. The collection of heritage documents can be found in the form of the oldest inscription until the modern printed books from the modern printing era in the 20th century. Although many major museums in the world have been trying to collect and to record the most important collections ever found, it is always very difficult to estimate the exact number of documents for some other older collections. Major projects in the world that are trying to do so must be supported by enormous resources with regard to human resources, time, and funds and other materials.

In practice, collaboration is often required between several parties from various fields of study incorporated in a consortium. The work to collect this collection of heritage documents must not only be supported by scientists and researchers, but the participation from public, government and industry is also absolutely necessary. This led to the high importance impact of every project of heritage document collections. In the last two decades, the efforts to collect heritage document collections began to be aggressively undertaken in various parts of the world. To accelerate the process of collecting a large number of collection of heritage documents, the tendency of the need for tools, systems, and automation procedures in the main process ultimately arises. Automated content extraction of heritage documents is a challenging task and hence an open issue for the re-

[3]http://www.historyofpaper.net/

search community [27]. More detailed description about the need of document analysis system for the collection of heritage documents will be given in Sub Section 2.1.3 of this dissertation.

Secondly, the physical condition of material of the document. Although materials such as clay plates and tablets, metal plates, papyrus and also paper are store-able for a long time, no material can survive forever against natural conditions over time. It is a common situation and condition in every part of the world that the collection of heritage documents found is not in perfect condition, even some of them are not in a readable condition anymore. As time goes by, there are of course various things that cause that condition to happen. The first thing is the natural causes that we cannot avoid, such as weather and climate conditions, natural disasters, and natural reactions involving physical, chemical, and biological factors in nature that tend to destroy material. The second thing is the human factor itself. There are many incidents that threaten and aggravate the condition of the collection of heritage documents, such as war, transformation of historic sites, and document storage that does not meet minimum standards to preserve the document.

Thirdly, the very valuable cultural content of the document. It is undeniable that there is an enormous variety and diversity of information contained in every collection of heritage documents. Important records concerning historical facts and the evolution of world civilization from the past are mostly contained in the written documents. Through these documents we can go back in time to see how civilization in every corner of the world began to take shape. A lot of answers to the essential questions that the world needs today are found in the collection of heritage documents. Among them even involves the sacred and secret information which provides us with a precious insight into older times beliefs and social standards. In some cases, they may be still relevant nowadays as a wisdom heritage for local communities. They are also helpful for consolidating the identity of a local community, its pride, and its unique features confronted to the cultural globalization. They can be applied very well as a local wisdom of certain communities in this modern era.

Fourthly, the possible access to the collection of documents, the need to share and to open the documents for the community and to make physical documents available to a large number of people [28]. Although many discoveries of the existence of heritage documents are widely known both in terms of quantity and in terms of the high value of the contents of the document, it does not necessarily mean all the access doors to the collection of documents is wide open. The access door intended here is not only the physical door of the building where the collection of documents is stored, but may be an official permit from the government if the collection concerns important historical documents of state, for ancient documents, the access door may also be local customary rules concerning the existence of the collection, the culture of the community around the location where the collection was found, along with the whole arrangement of beliefs that may be embraced by the people who own the document collection. In some cases, requesting permission to access and to share the collection of documents is not always easy. It happens sometimes that special circumstances make it impossible to publicly open the access to the collection of heritage documents and to share it with the wider community. Obstacles due to the inability to read and understand the contents of the document because the document uses language or scripts that are special or that have not been used in recent period also often occur. This case is widely experienced in projects for the document

collections with the scripts which were derived from non-Latin scripts, such as many heritage documents in the Asian region in general. For this case, the role of linguists and philologists is crucial in helping open the access to the contents of the document. Although this situation is often perceived as an obstacle in the implementation of projects for the heritage document, it is actually the condition that gives more added value to the project's implementation itself. Limited access to all the highly valuable content of the document collection is the main background and also as the principal goal and context of why such projects need to be done.

All four important points of interest in working with the collections of heritage documents described above can be clearly seen and be identified as real challenges as well as an integral part for the background and context of the research work in this dissertation. The palm leaf manuscript collections from Bali that became the main research object in this dissertation appeared exactly at the midpoint of those four points of interest.

2.1.3 The Need for a DIA System for Heritage Documents

Time, energy, and material cost are important matters when it comes to a project for a huge quantity of document collections. It is undeniable that several attempts to preserve and maintain the heritage document collections face the fundamental constraints of limited time, manpower and cost. It takes a considerable amount of time to be able to process thousands of document collections with a total of up to millions of pages, especially when all the processing steps must be done manually by human. Furthermore in addition, the preservation of these documents usually consist of a series of processes that are sequentially connected with each other. The pipeline is normally started with the physical document collection stage, and then cleaning the documents for further treatment, completed by the efforts to read, to transfer or to copy the contents of documents into other more recent media and format [29], and finalized by the efforts to disseminate the access to the contents of the documents to the wider public.

Since the advent of the digital world era in the 21st century, the conversion of physical documents into digital documents began to be campaigned for vigorously. Scanning physical document pages and storing them in a digital image format offers a greater chance of making the content of documents available to the digital world. Modern technology also provides a very strong support by providing digital data storage with a bigger capacity in smaller physical size and space. It also solves the problems of storage, paper deterioration, accessibility and many others [28]. From the last four decades, fast computers, large computer memory, and inexpensive scanners increase interest in digitizing physical documents into document image [30]. This makes the data format of the digital image to be the primary choice for physical document conversion. Normally as a digital data, the document image data format is presumably able to keep the contents of the document in a longer period of time. It no longer needs a large physical storage space, be easier and faster to be transferred, duplicated, and spread to the wider public.

However, with so many documents being digitized in this digitization campaign era, the volume of resulting document images is consequently also enormous. Although the effort to reduce the paper usage has not proved successful, the amount of digital document production is already too much [31]. Up to this critical point, it is finally considered the fact that the heritage document preservation efforts is not enough only to convert the physical documents into the document images. With many physical documents being

digitized and being stored in large document databases, and then being sent and received via digital machines, the interest and demand grew to provide more functionality with these images rather than to view and print them [30]. It needs further treatments until the collection of document images can be explored in more depth, fast and widespread. In simple examples, a more specific research field was needed to be developed to add machine capabilities in extracting information from these images, in reading text on a document page, finding sentences, and locating paragraphs, lines, words and symbols on a diagram [30].

To be able to accelerate the process of accessing, preserving and disseminating the contents of the heritage documents in a huge number of document images, a DIA system is needed. DIA means the process of using various technologies to extract text, printed or handwriting, and graphic from digitized document files[4]. DIA systems generally have a major role in identifying, analyzing, extracting, structuring, and transferring the document contents more quickly, effectively and efficiently. This system is able to work semi automatically or even more automatically without full human intervention. The DIA system is expected to accelerate time and to save cost and effort at many points in the heritage document preservation pipeline.

Technically, at the beginning of its emergence, the need for the DIA system comes from the basic functions of reading and extracting the contents of a document. For example, one of the basic functions in working with a document image is the character recognition function or more popularly known as the concept of Optical Character Recognition (OCR) [31, 32, 28]. This fundamental function arises when the result of digitizing the document into image format must be furthermore converted into text format. OCR is the core technology for a DIA software to extract text from images[4]. This is because the text formatting technology is already developed earlier and more advanced when imaging technology began to emerge. And for that reason, image document conversion to text format is intended to be faster and more adaptable for further processing such as text searching in multiple document collections or for text translation process from one language to another.

But over time, the number of digitized documents grew in terms of content formats, layouts, languages, alphabets, scripts, graphics, colors, and variations in the quality of the resulting document images. So, nowadays, the DIA system is required to perform more complex, established, advanced, and sophisticated functions. The DIA system is not only about OCR anymore. OCR is no more than a small component of much larger applications, and represents a small part of more complex system of what a DIA system really is [28]. Research to develop the DIA system with these new functions continues to grow rapidly and becomes very popular among computer science researchers in the field of digital image processing, pattern recognition, and in the last decade is closely engaged in research in the field of machine learning.

However, although the DIA research develops rapidly, it is undeniable that most of the document collections used as the research object in the initial step are mostly the documents, whether printed or hand written on paper, from the developed regions such as America and European countries. The document samples that come from these countries are mostly written in English or old English with Latin/Roman script. It was even

[4]http://www.cvisiontech.com/library/pdf/pdf-document/document-image-analysis.html

considered that English is blessed with one of the simplest scripts in the world [31]. Several important document collections were finally used as standard benchmarks for the evaluation of the latest DIA research results. The next wave of DIA research finally began to touch the documents from non-English speaking areas with non-Latin scripts, such as Arabic documents, Chinese, and Japanese documents. During the evolution of DIA research in the last two decades, DIA researchers have been able to propose and to achieve very satisfactory solutions for many complex problems of document analysis for these types of documents. But the DIA research challenge has not reached its end at that point. The latest challenge was coming when the documents from Asia with more new languages and more complex scripts were explored. For example Devanagari script [33], Gurmukhi script [34, 35, 36, 37], Bangla script [38], and Malayalam script [39], and in the case of multi languages and multi scripts documents from India. OCR for Indian languages in general is considered more difficult than for European languages because of the large number of vowels, consonants, and conjuncts (combination of vowels and consonants) [32].

This dissertation took part in exploring DIA research for palm leaf manuscript collections as the heritage documents from Southeast Asia, especially from Bali, Indonesia. This collection offers a new challenge for DIA researchers because it uses palm leaf as a writing media and also with a language and script that have never been analyzed before. A more detailed overview of DIA will be given in Chapter 4 of this dissertation.

2.2 Palm Leaf Manuscripts in Southeast Asia

Ancient manuscripts record much important knowledge about world civilization history. In Southeast Asia, most of the ancient manuscripts are written on palm leaf. Southeast Asia is a home for many ancient manuscripts where most of those manuscripts were handwriting on the dried palm leaves with complex languages and scripts. Ancient palm leaf manuscripts store various forms of knowledge and historical records including art, religion, and local wisdom from a long time ago. Many palm leaf manuscripts contain information on important issues such as medicine and village regulations that are used as daily guidance of social life in Southeast Asia. The collection of palm leaf manuscripts is one of the very valuable cultural heritages found in Southeast Asia.

For example in Cambodia, palm leaves have been used as a writing material dating back to the first appearance of Buddhism in the country. This type of manuscript is still seen in Buddhist establishments and is being used habitually and traditionally by monks to read scriptures. The languages written on the palm leaf documents vary from Khmer (the official language Cambodian people speak nowadays with slightly different spelling vocabularies) to Pali and Sanskrit by which the modern Khmer language is considerably influenced.

In Thailand, dried palm leaves have also been used as one of the most popular written documents for over five hundred years [21]. Such materials have been used for recording Buddhist teaching and doctrines, folklore, knowledge and use of herbal medicines, stories of dynasties, traditional arts and architectures, astrology, astronomy, and techniques of traditional massage.

Palm leaves were also historically used as writing supports in manuscripts from the

Indonesian archipelago. The leaves of sugar, or toddy, palm (*Borassus flabellifer*) are known as lontar. In Indonesia, the lontar-leaf manuscripts make up the vast majority of all known palm leaf manuscripts from Lombok to Sumatera [40]. Lontar are inscribed through a process of scratching or incising. Although the official language of Indonesia, Bahasa Indonesia, is written, nowadays, in the Latin script, Indonesia has many local, traditional scripts, most of which are ultimately derived from Brahmi [41]. In Indonesia, palm leaf manuscripts with different languages and scripts can be found in many regions. Although people still speak with their local languages, most of those scripts are not used anymore in their daily writing activities, for example, the Balinese script [15], the Javanese script, and the Sundanese script [42]. Nowadays they write using Latin script. Consequently, few people can still read and write with those scripts at this time.

The existence of ancient palm leaf manuscripts in Southeast Asia is very important both in term of quantity and variety of historical contents. It attracts historians, philologists, and archaeologists to discover more about the ancient ways of life. But unfortunately, the physical condition of natural materials from palm leaves certainly cannot last long and certainly cannot fight against time. Usually, palm leaf manuscripts are of poor quality since the documents have degraded over time due to age and due to inadequate storage conditions. They easily crumble under the attack of insects, fungi, and a humid tropical climate [43]. Many discovered palm leaf manuscripts in Southeast Asia are collections belongings to museums and private families that have been in a state of disrepair. Equipment that can be used to protect the palm leaf to prevent rapid deterioration are still relatively few in number.

2.3 Balinese Palm Leaf Manuscripts

Bali has a great social and cultural history with a rich tradition of literature that dates back several hundred years with the island's literary works mostly recorded on dried and treated palm leaves. The dried and treated palm leaf manuscripts in Bali are called lontar. For over a millennium, lontar were the medium for the transmission of knowledge both of a sacred or mundane nature; they are traditionally regarded as powerful, almost supernatural objects [44].

In Bali, palm leaf manuscripts were written in Balinese script in the Balinese language, in the ancient literary texts composed in the old Javanese language of Kawi and Sanskrit. Bali and Java have very close historical and cultural links for thousands of years. When Java received the influence of Islam, Bali maintained its Hindu culture. Therefore, the literary works in Bali hold much historical and cultural information from old Javanese culture. For example, Creese stated that there is a significant number of palm leaf manuscripts written in Balinese script which record law codes of Old Javanese [45].

The Balinese language is a Malayo-Polynesian language spoken by more than 3 million people mainly in Bali, Indonesia[5]. The Balinese language is the native language of the people of Bali, known locally as Basa Bali [41]. The Balinese language used today has gained much influence from many other languages such as Sanskrit, Old Javanese, Old Balinese, Modern Balinese, Sasak, Malay and Indonesian and some Dutch [46].

[5] www.omniglot.com/writing/balinese.htm

Figure 2.1: Palm leaf manuscript collections in the Museum of Gedong Kirtya, Singaraja, Bali

2.3.1 Collection from Bali, Indonesia

One of the special museums that holds the collection of Balinese palm leaf manuscripts is the Museum of Gedong Kirtya. The Museum of Gedong Kirtya is located in Singaraja, Regency of Buleleng, northern Bali. The museum was founded on 2 June 1928 by the Dutch in cooperation with Balinese advisers, under the name of Lontar Foundation Kirtya Liefrinck-Van der Tuuk. F.A. Liefrinck (1853-1927) and Herman Neubronner van der Tuuk (1824-1894) were two most popular Bali scholars in that period [46]. Herman Neubronner van der Tuuk is a colonial linguist who lived in Bali between 1870 and 1894. The Van der Tuuk collection of Balinese palm leaf manuscripts in the Leiden University library is the major source of nineteenth-century Balinese manuscripts [45, 47].

The opening ceremony of the museum was held on 14 September 1928 by Governor General Andries Cornelis Dirk de Graeff. Based on the Balinese çaka calendar, it was opened in the year of 1850[6]. The purpose of this foundation is to collect, to preseve, to make inventory and also to make copies of the important collections of palm leaf manuscripts.

In the catalog of the collection issued in 2015, there are 1757 titles/collections of palm leaf manuscript stored in this museum, with 800 collections already have copies, while 957 collections do not yet have a copy[7](Figure 2.1). But, between December 1928 and July 1941, the Foundation collected 2,263 lontar manuscripts [46]. It was also reported that previously there were about 2,111 titles/collections, but eventually some of the collections were damaged due to age[8]. One collection can consist of 5 or even up to 100 pages. In the Museum of Gedong Kirtya, the palm leaf manuscripts were stored in a wooden box as shown on figure 2.2.

Christiaan Hoykaas is a Dutch philologist who started the project of romanization of the Balinese palm leaf manuscript transcription, in 1939, together with I Gusti Ngurah Ketut Sangka. The Hooykaas-Ketut Sangka (HKS Project) Balinese Manuscript Collection can be found in the Australian National University Library[9]. The collection comprises approximately 6000 transcriptions. This project was resumed in 1972, and then continued by Hedi Hinzler and I Dewa Gede Catra in the 1980s until the first decade of the twenty-

[6]Bali Post, 24 May 2015

[7]Buku Katalog Lontar (Kropak), Pemerintah Kabupaten Buleleng, Dinas Kebudayaan dan Pariwisata Kabupaten Buleleng, U.P.T.D. Gedong Kirtya, Tahun 2015

[8]Bali Post, 24 May 2015

[9]https://trove.nla.gov.au/work/7638320?q&versionId=8799675

Figure 2.2: A wooden box to store the palm leaf manuscripts in the Museum of Gedong Kirtya, Singaraja, Bali

first century [43]. H. Hinzler was conducting field research visits to the northern, central, and southern parts of Bali around 1972 to 1992, and he published his findings about Balinese palm leaf manuscripts and the lontar palm tree in 1993 [46]. In his article, Hinzler compiled all the information that she collected with other main sources which date from the end of the nineteenth or the beginning of the twentieth century, such as the Balinese palm leaf manuscript collection from the Leiden University Library and one article by Grader and Hooykaas "Lontar als schrijfmateriaal" in 1941.

Some collections of palm leaf manuscripts from Lombok are also stored in the Museum of Gedong Kirtya. The epic of lontar varies from ordinary texts to Bali's most sacred writings. They include texts on religion, holy formulae, rituals, family genealogies, law codes, treaties on medicine (usadha), arts and architecture, calendar, prose, poem and even magic. In his article, Hinzler mentioned that royal families in North Bali (Buleleng) and South Bali (in the regencies of Karangasem, Klungkung, Bangli, Badung and Tabanan), such as kings, princes and princesses also wrote poetry [46]. With a great influence from Indian culture, the Balinese manuscript's content were mostly based on the famous Indian epics of Ramayana and Mahabharata.

Other collections of palm leaf manuscripts can be found in the libraries of Udayana University. Around 3,000 collections are in the Pusat Dokumentasi Dinas Kebudayaan Provinsi Bali[10] (Centre for Documentation of Balinese Culture), in the Museum Bali in the City of Denpasar. Museum Bali also stores many important collections of palm leaf manuscripts. Apart from the collection at the museum (Museum Gedong Kirtya Singaraja and Museum Bali Denpasar), it was estimated that there are more than 50,000 lontar collections which are owned by private families (Figure 2.3). Unfortunately, in reality, the majority of Balinese have never read any lontar because of language obstacles as well as tradition which perceived them as a sacrilege[11]. In Museum Bali, we can find some transliteration (in the Indonesian language : ***Alih Aksara***) books of the manuscript into the Latin/Roman alphabet (Figure 2.4), but it is still hard to associate the collection and their transliteration book. In some cases, a transliteration book belongs to a collection which is not stored in that Museum.

[10] http://bali.tribunnews.com/2014/09/12/melihat-dari-dekat-pusat-dokumentasi-dinas-kebudayaan-pemprov-bali

[11] January 29, 2011 (The Jakarta Post)

Figure 2.3: Private family collections in Village of Jagaraga

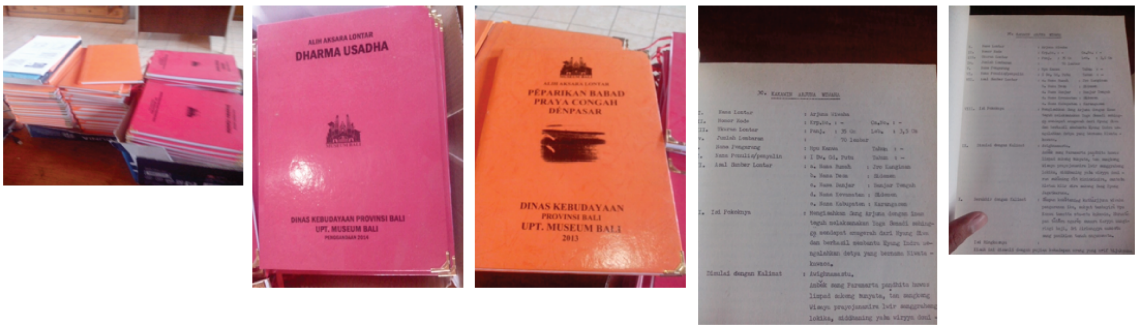


Figure 2.4: Some transliteration (in Indonesian language : ***Alih Aksara***) books of the manuscript into the Latin/Roman alphabet in Museum Bali, Denpasar

In 2011, the Internet Archive[12] digitized (by digital photography)[13][14], uploaded and published around 2,762 Balinese palm leaf manuscript collections from the major library in Denpasar, Bali. Most of the collection come from the Centre for Documentation of Balinese Culture in Denpasar [43].

2.3.2 Processing palm leaves into writing material

Hinzler stated that there is not much written information that can be found about the Balinese palm leaf manuscripts and their production process around the nineteenth and early twentieth century. And there are not many references on the production process of palm leaf as writing material before the nineteenth century.

Hinzler [46] reported that the number of specialists processing the leaves was small. This is in accordance with the reality that until now in Bali, the main demand for the palm leaves from the Balinese people is actually for the material to make offerings at religious ceremonies. For this task, the palm leaves do not require the same processing as the palm leaves used for writing material. To make the offering, palm leaves in dry condition are sufficient. For the larger needs, young green coconut leaves serve the purpose, although it can not last as long as dried palm leaves. That is why until now, there are not so many Balinese people who know the details of the production process of palm leaf as writing material.

[12]https://archive.org/details/Bali&tab=collection
[13]https://balebengong.id/uncategorized-id/menjaga-lontar-bali-agar-tak-telantar.html?lang=id
[14]January 29, 2011 (The Jakarta Post)

Figure 2.5: Palm tree

Notwithstanding paper has become the main writing material used by modern Balinese society, the production of palm leaves into a writing material still exists and can be found in several places in Bali. This small scale production of palm leaves is primarily intended to fill the demand for the writing tasks of religious documents for certain ceremonies. However, it is now also popularly used for the production of souvenir objects for the tourists in the form of writing samples on the new artistically decorated palm leaves.

Hinzler found that Grader mentioned trees concentrated in the Regency of Karangasem as sources of best quality leaves. As suggested by one Balinese philologist who works for the AMADI project[15], we did an observation research visit to the Village of Sidemen, Regency of Karangasem, Bali, to observe and to understand the detailed process of palm leaf production as writing material. We met with one of the residents in the village who still produces palm leaves for writing material from generation to generation in his family. We then compared our findings to the reported work of Hinzler who captured the situation and condition from fourty five years ago around 1972. Based on our observation, the palm leaf production process into writing material is in general still done with the same methods now.

The selected palm tree is from the species of *Borassus flabellifer oxflabelliformis*, or *Palmyra* (Figure 2.5).

Normally, the palm trees have fan-shaped leaves. The best palm leaves should be old

[15]Personal communication: Bapak I Made Muliarta, City of Singaraja, 2015

Figure 2.6: Selection of palm leaves

Figure 2.7: Cutting the palm leaves

enough, wide, smooth and flat. The width and length of palm leaves are also important considerations in choosing the palm leaves that can be used as a good writing material. The average dimension of the raw palm leaves are 1 m long and 3 cm (at the tip) to 6 cm (at the base) wide (Figure 2.6).

Before it can be used as writing material, natural palm leaves must be processed first. The purpose of natural palm leaves processing is to produce dried palm leaves that can last physically for long periods of time, and is expected to be protected from the threat of damage by certain insects.

First, the palm leaves are cut into the desired length. The selected part is usually the widest part at the base of the palm leaf. Palm leaves with a long and wide size are usually used to write the most important collections. After cutting, the ribs of the palm leaves are removed (Figure 2.7). Subsequently, palm leaves should be soaked in water to become softer. This soaking process is usually done for approximately ten days[16]. The tool used to soak the palm leaves is also very simple, simply a plastic bucket (Figure 2.8). The same technique, tool and method are also described in Hinzler's article [46].

Palm leaves are then boiled with water and some spices for about 4 hours (Figure 2.9). At the time of our observation, the spices used were tea and pepper. Hinzler mentioned in his article that there are various recipes for this herbed bath. Some spices and herbs may be used for this process, for example bark of the *intaran* tree (*Azadirachta indica*), bark of the book tree (*Dracontomelon*), root of the *sikapa* tree (*Dioscorea hispida*), the bud of the coconut and the *temitis* plant (*Curcuma xanthorrhiza*), leaves of the *liliguncli* (*Vitex trifolia*), leaves and shoots of the *Uncarica gambir* (not too many, otherwise the leaves are coloured too red) and *kunyit warangan* (*Curcuma longa*), *tengeh*, the reddish bark from the *tingi*

[16]Personal communication: Bapak Dewa, Village of Sidemen, 2015

Figure 2.8: Soaking the palm leaves

Figure 2.9: Boiling the palm leaves with water, spices and herbs

tree (*Bruguiera parviflora*), *bejadi*, a kind of *santen* (coconut milk) from which the oil has been removed, the bark from the intaran tree, and *palapah borèh*, an ointment consisting of *kesuna* (garlic), *isèn* (*langkwas* root) or *isin rong* (*Alpinia galanga*) [46]. To the best of our knowledge, the names of some spices and herbs are sometimes different across regions in Bali. Some plants are difficult to be associated with plants that still exist today, or perhaps the plants are already getting rare. These spices and herbs will make palm leaves resistant from attack by insects.

After the boiling process with spices and herbs, palm leaves are then dried under the sun (Figure 2.10). This process will dry out the leaves and change the texture to be slightly wrinkly and wavy. (Figure 2.11). To remove wrinkles and wavy surfaces from this palm leaf, palm leaves should be clamped using two wooden blades (Figure 2.12). The longer the palm leaves are clamped, the better the quality of palm leaves produced for the writing material. Two wooden blades are also used at once to smooth the size of the palm leaves. The size of flat palm leaves are adjusted again using a small cutter like knives and the wooden slats as the base size (Figure 2.13).

The next step is to make holes in the palm leaves. Normally, there are three holes on a sheet of palm leaf. A hole on the left and right section, as well as a hole positioned almost in the middle of the leaf (Figure 2.14). The size of the hole is quite small, only about 3 to 5 mm in diameter. The holes on the left and right sections of the leaf are about 2 cm from the edge of the leaf, while the hole in the middle of the leaf is not exactly in the center of the leaf. According to Bapak Dewa[17], the position of the middle hole will determine the

[17]Personal communication: Bapak Dewa, Village of Sidemen, 2015

Figure 2.10: Drying the palm leaves under the sun

Figure 2.11: Dried, wrinkle and wavy palm leaves

Figure 2.12: Wooden blades for clamping palm leaves

Figure 2.13: Adjusting the size of the palm leaves

balance between the left and the right side of the leaf and it will be used to determine the starting point of writing from the right side. A wooden mold is used to determine the position of the three holes (Figure 2.15). The positions are then marked by using charcoal or using a pencil (Figure 2.16). For punching the holes, a needle-like instrument is used. The hole is made by turning the instrument around (Figure 2.17). After making the three holes, the palm leaves should be clamped for the second time with two wooden blades. Bamboo blades are also pinned through the three holes from a bunch of palm leaves (Figure 2.18). The edges of the palm leaves are then smoothed with pumice or sandpaper (Figure 2.19). Sometimes, on some palm leaf manuscript collections, the edges of the palm leaves are colored red (Figure 2.20). Hinzler also mentioned that there are quite a few exquisite red colored materials imported from China to color important collections [46]. Some collections may be coloured with gold or red with ornaments (Figure 2.21).

After all of those processes, the palm leaves are now ready to be used for writing. To facilitate the process of writing, the helping horizontal lines will be provided on the on palm leaves (Figure 2.22). Yarn or thread and charcoal are used to draw the horizontal lines on the surface of palm leaves (Figure 2.23). Yarn or thread are dyed with a mixture of charcoal and oil, and then it is pressed on the surface of the leaf to make a trace of black horizontal lines (Figure 2.24). But in the modern era, ruler and pencil can be used to draw those lines (Figure 2.25). The lines should not be too thick or too visible so it can be easily removed after the writing process. In most of the palm leaf manuscript collections, one page usually contains four text lines (Figure 2.26). Some of the collections can also contain only three text lines (Figure 2.27) or with five text lines per page (Figure 2.28). For some special needs in writing format, for example a table (Figure 2.29), some vertical lines are also provided to facilitate the writing process. The vertical lines can be also used as a margin of the page (Figure 2.30). After the palm leaves are inscribed, in some private

Figure 2.14: Three holes on a palm leaf

Figure 2.15: Determining the position of the three holes on the leaf

Figure 2.16: Charcoal or pencil to mark the hole position on palm leaves

Figure 2.17: Punching the holes on palm leaves

Figure 2.18: Clamping and pinning the palm leaves

Figure 2.19: Smoothing the edges of palm leaves with sandpaper

Figure 2.20: Red coloring on the edges of palm leaves

Figure 2.21: Ornaments on palm leaf manuscripts

Figure 2.22: Horizontal lines on palm leaves to facilitate the writing process

family collections, the manuscript is kept at the top of the traditional kitchen fireplace. Smoke from the fireplace will protect the palm leaves from insect attack, but it makes the palm leaves look darker, harder and stiffer. For example, a darker manuscript collection in AMADI_LontarSet is coming from a family in the Village of Rendang, Regency of Karangasem (Figure 2.31).

2.3.3 How to write a Palm Leaf Manuscript

To write on lontar leaves, a tool like a small pen-knife called *pengerupak* is needed. The *Pengerupak* is made of iron (Figure 2.32). The text is written by scratching the palm leaves with a *pengerupak*. Lontar leaves that have been scratched will be rubbed with natural black dye for example by using burned candlenut to produce a black color like carbon. And then, the lontar leaves will be cleaned up with cotton, and the black color will stay on the scratched part as text (Figure 2.33). The Balinese palm leaf manuscripts can contain

Figure 2.23: Yarn and charcoal are used to draw the straight horizontal lines

Figure 2.24: Drawing the horizontal lines on palm leaves

Figure 2.25: Ruler is used in modern era to draw the lines

Figure 2.26: A page of manuscript with four text lines: MB-AdiParwa(Purana)-5338.2-IV.a-P3.jpg

Figure 2.27: A page of manuscript with three text lines: JG-02-P7.jpg

Figure 2.28: A page of manuscript with five text lines: Bangli-P59.jpg

Figure 2.29: A page of manuscript with vertical lines used as a table: WN-P11a.jpg

Figure 2.30: A page of manuscript with vertical lines used as a margin: Bangli-P54.jpg

Figure 2.31: Darker, harder and stiffer palm leaves from the Village of Rendang

Figure 2.32: Pengerupak

only text, only graphics, or both text and graphics.

2.3.4 Balinese Language and Balinese Script

The Balinese script, or ***Aksara Bali***, is used for writing the Balinese language. The Balinese script is a descendant of the ancient Brahmi script, and without doubt, is derived from the Devanagari and Pallava scripts from India. Therefore, it has many similarities with modern scripts of South Asia and Southeast Asia, which are also members of that family. The shape of the script shows similarities with southern Indian scripts like Tamil. The concept of syllable is also found in other South/Southeast Asian scripts, such as the modern Devanagari, Tamil, Thai, Lao, and Khmer scripts. The Balinese script's closest sibling is the Javanese script, which have rectangular form of font shape compared to the round shape of the Balinese script [48]. Balinese and Javanese are closely related, highly ornate scripts; Balinese is used for the Balinese language on the island of Bali, and Javanese for the Javanese language on the island of Java [41]. The Balinese script is used to write Kawi, or Old Javanese, which strongly influenced the Balinese language in the eleventh century. In history, the Balinese script in Bali has not been replaced by the Pegon script which is an Arabic alphabet used to write the Javanese and Sundanese languages, like in West Java and in Malaysia.

Figure 2.33: Writing on palm leaves

The Balinese script is considered to be one of the more complex scripts from Southeast Asia. The type of writing system is syllabic alphabet. In general, in Balinese script[18]:

- the vowel "A" is implicit after all consonants and consonant clusters and should be supplied in transliteration, unless: (a) another vowel is indicated by the appropriate sign; or (b) the absence of any vowel is indicated by the use of an "ADEG-ADEG" sign.
- vowels are almost always indicated by one of a class of agglutinating signs (Pangangge Suara) added above, below, before, or after the consonant or consonant cluster which they affect.

According to the Unicode Standard 9.0[19], Balinese script has the Unicode table from 1B00 to 1B7F (see Apendix B), but a complete set of Balinese script contains **156 glyphs** as follow:

- **121 glyphs are listed in the Unicode Table** (Appendix B). It consists:
  - **33 Consonants**, comprising:
    - **18 "Basic" Consonants** (Appendix C Table C.1). The Balinese script is popularly known as "***HANACARAKA***". It represents the first five syllables of these basic consonants: ***HA***, ***NA***, ***CA***, ***RA***, and ***KA***).
    - **15 "Special" Consonants** (Appendix C Table C.3). These specific consonants represent some specific speech sound of syllables which often come from Sanskrit words.
  - **14 Vowels / Independent Vowels** consist of:
    - **7 "Basic" Vowels** (Appendix C Table C.5)
    - **7 "Longer" Vowels** (Appendix C Table C.6). These vowels represent the "longer" version speech sound of syllables of basic vowels.

[18]https://www.loc.gov/catdir/cpso/romanization/balinese.pdf
[19]http://www.unicode.org/versions/Unicode9.0.0/

  - **18 Pangangge**. It consists:
    - **14 Pangangge Suara / Dependent Vowels** (Appendix C Table C.7)
    - **4 Pangangge Tengenan** (Appendix C Table C.9)
    - Actually there is another type of Pangangge, called **Pangangge Aksara**. It consists 3 glyphs to represent the conjunct forms of 3 basic consonants, so they are not listed in the Unicode Table (Appendix C Table C.10).
  - **10 Digits** (Appendix C Table C.11)
  - **8 Punctuations** (Appendix C Table C.13)
  - **3 Additional Signs** (Appendix C Table C.12)
  - **28 Musical Symbols**
  - **7 Additional Consonants for Sasak**

- **35 glyphs not listed in the Unicode Table** consist of:
  - **30 Conjunct Forms of Consonants (usually called *Gantungan* or *Gempelan*)** comprising:
    - **15 Conjunct Forms of "Basic" Consonants** (Appendix C Table C.2). It should be 18 Conjunct Forms of Basic Consonants, but as explained above, 3 more glyphs of conjunct forms of basic consonants are categorized as **Pangangge Aksara** (Appendix C Table C.10).
    - **15 Conjunct Forms of "Special" Consonants** (Appendix C Table C.4)
  - **3 Pangangge Aksara** (Appendix C Table C.10)
  - **2 Additional Symbols** (Appendix C Table C.12)

Some glyphs are written on the upper baseline (Ascender) or under the baseline of the text line (Descender). Figure 2.34 shows three different examples of how to write ***wanada***, ***winadu***, and ***windu*** in Balinese script. These examples show how to use *Pangangge Suara* and Conjunct Forms of Consonants.

2.4 Conclusions

This chapter provides a comprehensive overview of the social, cultural and historical aspects of palm leaf manuscripts collection in Southeast Asia in general, and in particular it describes the existing situation and condition about palm leaf manuscripts collection in Bali. Palm leaf manuscripts represent Southeast Asian region in the historical timeline of the invention of writing materials. In Bali, Indonesia, palm leaf manuscripts can be considered as the principal literary heritage which records much valuable knowledge of the Balinese people and society. Although it is quite difficult to estimate the exact number of the entire lontar collection in Bali, the information on the collection in the main museums can at least give an idea of the importance quantity of some of the collections. The collecting periods can also be known from the information provided based on the catalog of the Herman Neubronner van der Tuuk collection. This chapter describes in detail the steps to process the natural palm leaves into palm leaves that will be used as a writing material. Based on the observation to the field, it can be seen that the methods and tools used to process palm leaves are quite simple and people still do the same as what was done about fourty five years ago based on the reported work of Hinzler [46]. This is one proof of how strong the social aspect of palm leaf exists in Bali. This proves

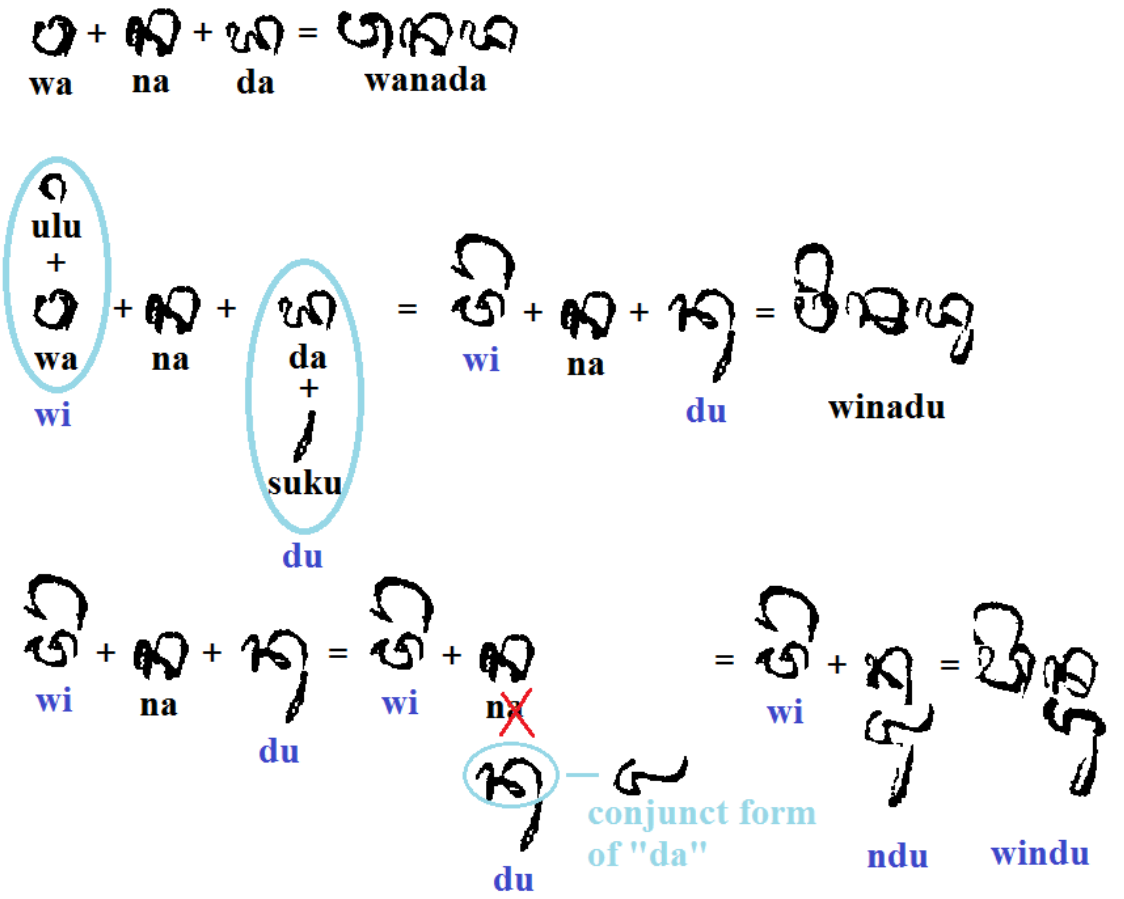


Figure 2.34: An example on how to write *"windu"* in Balinese script, with the different examples on how to write ***wanada*** and ***winadu***. The syllable *"**wi**"* is composed by two glyphs while the syllable *"**ndu**"* is composed by three glyphs. The glyph combinations are mostly written in vertical position. In this example, five images of glyph in two units of vertical arrangement, should be mapped (transliterated) into five Latin letters in horizontal arrangement.

that the writing culture with palm leaves is strongly attached to the community in Bali.

With a huge quantity including the private families collection, but on the other hand with a very limited access to their content, the existence of the palm leaf manuscript collections provides a great and complete interest towards the development of document image analysis system. The fact that Balinese palm leaf manuscripts were written in a complex syllabic script with a large glyph set, clearly shows the need of a transliteration system. To develop this system, a multidisciplinary collaboration between computer scientists, linguists and philologists is absolutely needed. In the next chapter, the challenges in developing the DIA system for Balinese palm leaf manuscripts will be described.

Chapter 3

Challenges for developing a DIA System for Balinese Palm Leaf Manuscripts

This chapter describes and discusses the challenges for the development of a DIA system for Balinese palm leaf manuscripts in three parts. The first challenge is the socio-cultural aspects. The second challenge is the physical characteristic of the manuscripts, and the third challenge is the complexity of the Balinese script.

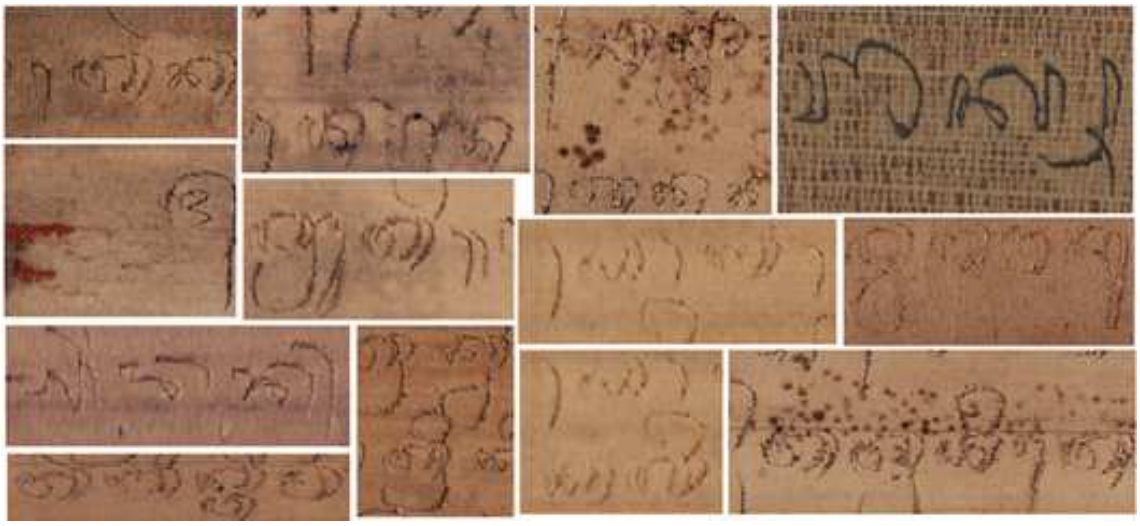


Figure 3.1: The degradations on palm leaf manuscripts [1]

There are several major challenges in working with the collection of palm leaf manuscripts. Those challenges are not only about technical challenges, but also socio-cultural challenges as a point of interest associated with the collection of Balinese palm leaf manuscripts. The technical challenges for palm leaf manuscripts in DIA system are twofold:

- The first challenge is the physical condition of the palm leaf manuscripts which will strongly influence the quality of the captured document images (Figure 3.1). For the image capturing process for DIA research, data in a paper document are usually captured by optical scanning, but when the document is on a different medium such as microfilm, palm leaves, or fabric, photographic methods are often used to capture images [32]. And nowadays, due to the specific characteristics of the physical support of the manuscript, the development of DIA methods for palm leaf manuscripts in order to extract relevant information is considered as a challenging problem in handwritten document analysis [21, 49, 50, 23, 14, 1, 25, 18, 15, 11]. It ranges from the binarization process [1, 18, 23], text line segmentation [25], character recognition and text transliteration tasks [23, 14] to the word spotting methods.

- The second challenge is the complexity of the Balinese script. The Balinese palm leaf manuscripts with different scripts and languages surely provide some real challenges for document analysis methods, not only because the different forms of characters from the script, but also the writing style for each script differs in how to write and to join or separate a character in a text line.

3.1 Socio-cultural Challenges

The first socio-cultural challenge is the difficulty in collecting sample manuscripts to create the initial corpus. The Balinese manuscript collection is spread over the two main museums owned by the local government of Bali, but the largest manuscript collection is the private collection owned by each Balinese families. Although the number of collections is widely assumed, the manuscript digitization campaign is often inhibited because there are cultural and religious conditions that cannot be ignored and passed away. Some

of the collections are still considered sacred which means it is not so easy to get permission to open, to view, and to digitize them. The access to the collections in the museums is also limited to only a small part of the whole collection of the manuscripts. There are also some rules to be obeyed during the capturing process for the manuscripts. The sacred manuscripts can not be carried anywhere carelessly. In Bali, many more manuscripts are known to be stored by families, who are very reluctant to lend it because of the supposed magic power of these old and respected artefacts. A safe and efficient standard procedure for digitizing was designed, in order to preserve the physical integrity of the manuscripts and to ensure a necessary quality to analyse the document. Due to the fragility of some documents, classical scanner devices can not be used.

The second socio-cultural challenge is the difficulty in finding the Balinese philologist to work within this project. Annotating a complex script like the Balinese script in palm leaf manuscripts require language specific expertise [29, 51].There are not many Balinese who can read well the Balinese script. Actually, the Balinese script is taught in elementary school for all students. However, since the Balinese script is not used in everyday writing, most Balinese do not remember how to write and read it after graduating school. They speak Balinese, but they write in Latin/Roman script. Moreover, it is commonly known that Balinese philology is not a popular field of study among the young Balinese generation. Most of the Balinese philologists are already in retirement. In one side, working with a younger philologist is actually a better choice as we know that a considerable amount of time and energy will be really needed to work in this project. The younger generation is indeed also more adaptable in using some digital tools or applications which will maybe used in the framework of this project. But, on the other side, these younger philologists still lack experience in analyzing ancient texts.

3.2 Physical Characteristic of Balinese Palm Leaf Manuscripts

Although palm leaves have been processed in such a way to prevent insects' attack and they are expected to survive for a longer period of time, it is quite often that the condition and location to store the collection of palm leaf collection is not good enough. Most of the collections found, especially those stored in private family homes, are already in a degraded state and disrepair. We identify several types of physical degradation found in the corpus of palm leaf manuscripts as follow.

- Low intensity variations, poor contrast (Figure 3.2). Natural dyes used to color the scratched part of the text on the palm leaves faded. This may be due to dye mixture used that is not strong enough such that it becomes faded easier. It can also be caused by the weak pressure of the *pengerupak* when scratching on the palm leaves, so the dye does not penetrate into the palm leaves strongly. It creates the fading gray levels of text and discoloured manuscripts (Figure 3.3).

- Other than natural aging cause, the humid storage conditions cause the appearance of mold, with foxing and yellowing area on the palm leaves. In private homes, the manuscripts are rarely opened for aeration. Storage in smoky areas like above the fireplace also provides black nuances on palm leaves (Figure 3.4).

- Random noise because of marks of strain from the palm leave texture can also be found on the background parts of the manuscripts (Figure 3.5).

Figure 3.2: Low intensity variations, poor contrast

Figure 3.3: Discoloured document, fading gray levels of text

- Additional damage to the resulting digital image may also come from an imperfect digitization process (Figure 3.6).

These physical degradations make the binarization task, the text line segmentation and glyph segmentation task quite challenging in the DIA pipeline for Balinese palm leaf manuscripts.

3.2.1 Challenges in Binarization

The binarization process, as one of the early and important stage in document analysis pipeline, is also a real challenge for ancient palm leaf manuscripts. Written on a dried palm leaf by using a sharp pen (which looks like a small knife) and colored with natural dyes, it is hard to separate the text from the background in the binarization process. With the aim of finding an optimal binarization method for palm leaf manuscripts, some binarization methods which have already been proposed and widely used in document image research community have been tested and evaluated. In our previous work [1], we have experimented and compared several alternative well-known binarization algorithms on our palm leaf manuscripts. Figure 3.7 shows the binarized images when applying different method such as Otsu [2, 3], Niblack [4, 5, 6, 7, 8], Sauvola [9], Wolf [4, 5], Rais [5], NICK [4], and Howe [10]. Based on a visual observation to compare the results, it is clear that those binarization methods do not give a good binarized image for palm leaf manuscripts. All methods extract unrecognizable characters on palm leaf manuscripts with noise. One other challenge is the need to create the binarized ground truth image. Since there is no existing ground truth binarized image for our palm leaf manuscripts, we cannot objectively evaluate these results. Therefore, to binarize the images of palm leaf manuscripts, a specific and adapted binarization technique is required.

Figure 3.4: Artifacts due to aging, black nuances, foxing, and yellowing

Figure 3.5: Random noise

Figure 3.6: Scanning errors and resolution, problems from the conversion to digital image

3.2.2 Challenges in Text Line and Glyph Segmentation

Even though some of the text line segmentation methods are already performed very well in a printed document, segmenting the text lines in a handwritten document is obviously challenging. In a handwritten document, the spatial positions of the words and the characters that compose the text lines are possibly not in a regular straight horizontal and vertical positions. They are often arranged in a skewed medial axis and in some documents they form a curved base line. These irregular conditions of the medial axis and base lines directly and greatly increase the challenge to detect the separating paths between the text lines in text line segmentation process. The palm leaf manuscripts contain some obstacles for line segmentation, for example skewed and fluctuating text lines, and irregularity in geometrical properties of the line, such as line width, height, and distance in between lines [52] (Figure 3.8). The variation in size of the characters and the different spaces between text lines further complicates the text line segmentation task such that characters may be too close to each other or oversized characters that cover two consecutive text lines exist. (Figure 3.9). For historical documents that were written in Asian-type scripts, the existence of many diacritics or other smaller sized characters that were written separately above or under the main text line is another real challenge for text line and glyph segmentation.

Naturally, Balinese script is not a cursive writing style. The glyphs are not written joined together so the text line and glyph segmentation is still quite reasonable to do. In Balinese script writing, there are no spaces between words from left to right in a horizontal text line. With no spaces between words, the next level of segmentation that must be done after the level of text line segmentation is isolated glyph segmentation as the individual image of symbol. The glyph segmentation also suffers the same variation of space between glyphs (Figure 3.10).

Most line and glyph segmentation approaches in the literature require binarized image as input. However, due to degradation and noise often found in historical documents such as palm leaf manuscripts, the binarization task, the text line and glyph segmentation methods are not able to produce decent results. In this case, a good initial binarization

Figure 3.7: Original image (upper left), and binarized images (top to bottom, left to right) [1] using methods of Otsu [2, 3], Niblack [4, 5, 6, 7, 8], Sauvola [9], Wolf [4, 5], Rais [5], NICK [4], and Howe [10]. More detailed experimental results are presented in Appendix D.

Figure 3.8: Varying space between lines (leading)

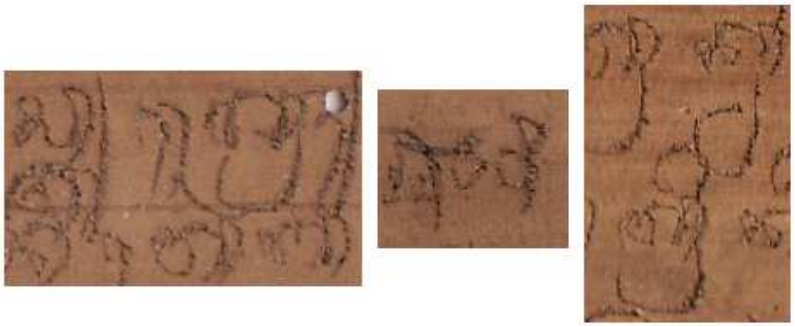


Figure 3.9: Merges, fractures and other deformations in the character shapes

Figure 3.10: Varying space between glyphs (kerning)

process is required. Unfortunately, as already mentioned in Sub Section 3.2.1, the binarization process to separate the ancient text from the background is a real challenge for some types of historical document image like the palm leaf manuscripts from Southeast Asia [1, 18, 22].

3.3 Complexity of Balinese Script

In the domain of DIA, the handwritten character recognition has been the subject of intensive research during the last three decades. Some methods have already reached a satisfactory performance especially for Latin, Chinese and Japanese script. However, the development of handwritten character recognition methods for other various Asian scripts, such as Devanagari script [33], Gurmukhi script [34, 35, 36, 37], Bangla script [38], and Malayalam script [39], always presents many issues.

In the OCR task and development for Balinese palm leaf manuscripts, severals deformations in the character shapes are visible due to the merges and fractures of the use of nonstandard fonts. The similarities of distinct character shapes, the overlaps, and interconnection of the neighboring characters further complicate the problem of OCR system [53] (Figure 3.11). One of the main problems faced when dealing with segmented handwritten character recognition is the ambiguity and illegibility of the characters [54]. These characteristics provide a suitable condition to test and evaluate the robustness of feature extraction methods which were already proposed for character recognition.

3.3.1 Challenges in Isolated Glyph Recognition

Just like binarization, text line segmentation and isolated glyph segmentation, isolated glyph recognition also suffers from physical degradation of the manuscripts. In addition, there are other challenges for glyph identifier or glyph recognizer which are influenced by the nature of the Balinese script itself.

- Like any other handwritten document, different styles in writing the glyphs of the

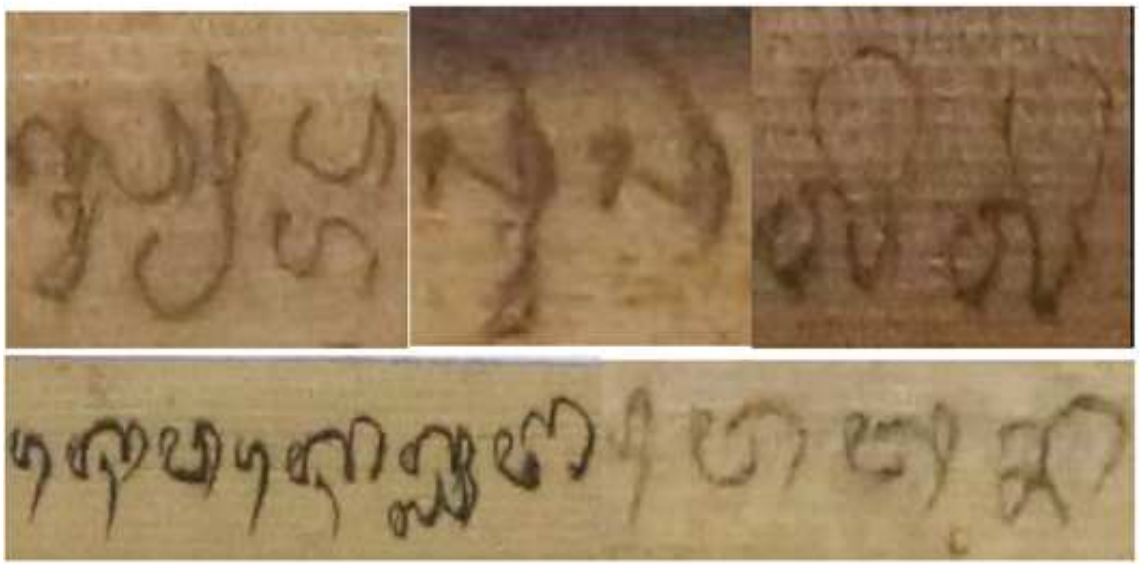


Figure 3.11: Balinese script on palm leaf manuscripts [11]

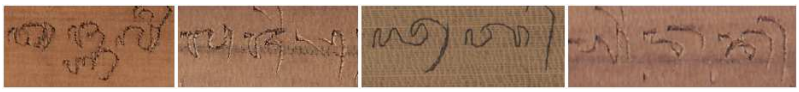


Figure 3.12: Different styles in writing Balinese script from different writers

Balinese script from different writers naturally exist (Figure 3.12). The length of strokes (or scratches in this type of manuscript) for a glyph often varies between writers. As a result, in some cases, the height or width of a glyph appears very different from the standard proportion size commonly found for that glyph. For example, the glyph "***TALING***", "***ADEG-ADEG***", "***BISAH***", and "***SUKU***" can have different length of strokes in vertical direction (Figure 3.13), and the glyph "***CE-CEK***" and "***PEPET***" can be represented with many different dimensions of strokes (Figure 3.14).

- The problem of interclass similarity. In many cases, the shape of one glyph is quite similar to another glyph and a part of glyph shape is exactly the same with another glyph shape. For example, a part of the shape of glyph "***KA***" share exactly the same shape as the glyph "***NA***", and the shape of glyph "***BA KEMBANG***" is quite similar to the shape of glyph "***DA***" (Figure 3.15).

3.3.2 Challenges in Text Transliteration

The challenge for the transliteration of the Balinese script comes from the fact that it is a syllabic script and that the speech sound of the syllable change related to some certain phonological rules [26]. The mapping between linguistic symbols and images of symbols is not straightforward (Figure 2.34). For the text transliteration, the challenges are identified as follow.

- The problem of one-to-one mapping between linguistic symbols and images of symbols [20]. The glyph agglutination can be found between consonant glyphs with the conjunct forms of other consonant glyphs or with the dependent vowel

Figure 3.13: Different height of glyphs *"TALING"*, *"ADEG-ADEG"*, *"BISAH"*, and *"SUKU"* from different writers

Figure 3.14: Different proportion size of glyphs *"CECEK"* and *"PEPET"* from different writers

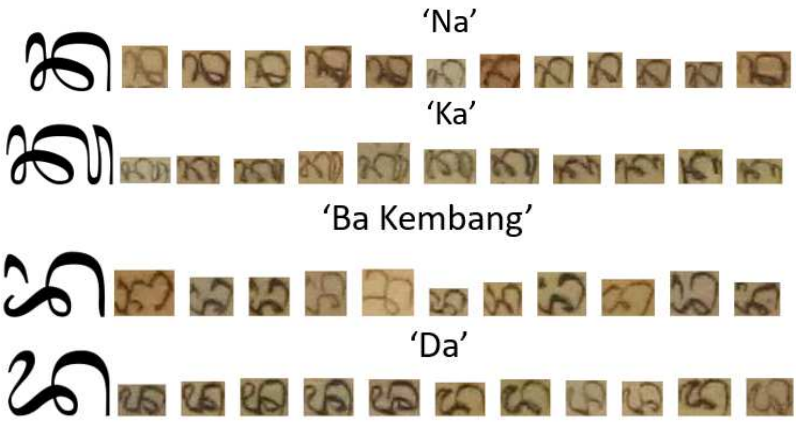


Figure 3.15: Interclass similarity between glyph *"NA"* vs *"KA"* and *"BA KEMBANG"* vs *"DA"*

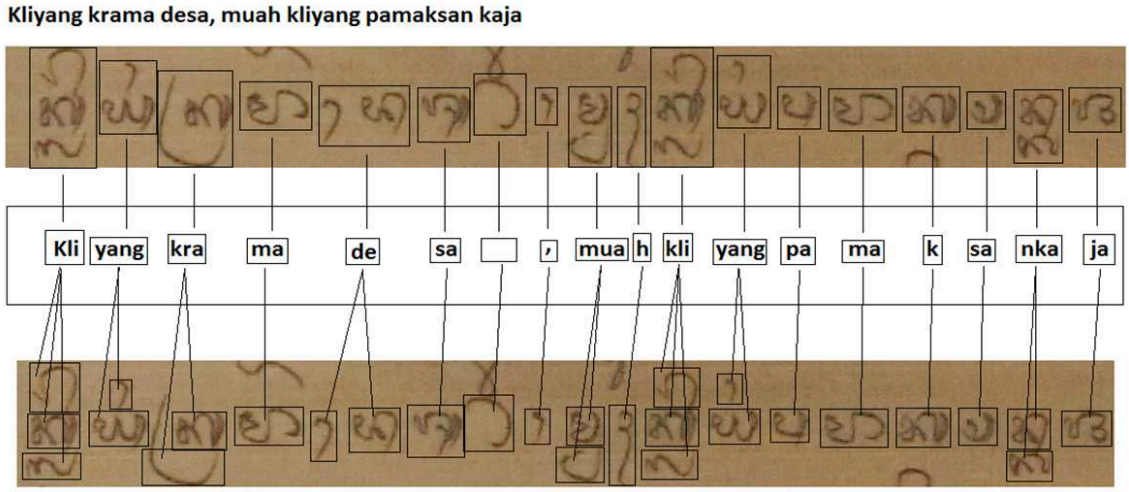


Figure 3.16: Huge combination of possible compound syllable

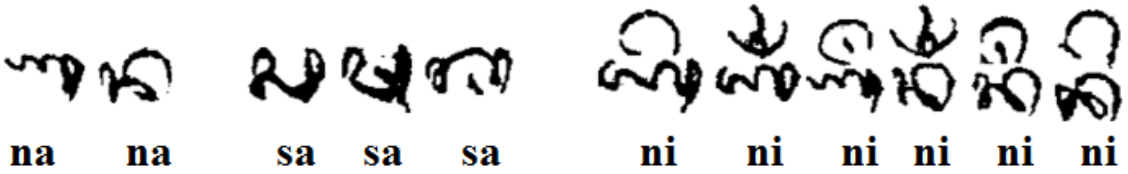


Figure 3.17: Allographs for syllable "*NA*" with two possible glyphs, "*SA*" with three possible glyphs, and "*NI*" with six possible combinations of glyph. But inconsistency are often found in transliterating these allographs, for example "*SA*" or "*SHA*" or "*SSA*"

glyphs. The agglutination in the vertical position makes the mapping between glyph symbols and their speech sound of syllable not exclusively one-to-one. One or more basic glyphs can join to produce a compound syllable, or one syllable can be mapped to one or more glyphs.

- Defining a compound syllable as the fundamental unit of writing is possible, but the number of combination of possible compound syllable will be huge (Figure 3.16), and collecting enough labeled samples for each class is hard and it needs an extraordinary effort.
- The problem of "*allographs*" [20], where more than one shape of glyph (image of symbol) is allowed to be used to represent a same sound of speech of syllable (linguistic symbol). For example, the syllable "*NA*" can be possibly written with two different glyphs, glyph "***NA***" or glyph "***NA RAMBAT***", the syllable "*SA*" can be possibly written with three different glyphs, glyph "***SA / SA DANTI***", glyph "***SA SAPA***", or glyph "***SA SAGA***", and the syllable "*NI*" can be possibly written in six different combinations of glyph "***NA***" or "***NA RAMBAT***", and glyph "***ULU***", "***ULU CANDRA***", or "***ULU SARI***" (Figure 3.17).

3.4 Conclusions

This chapter shows the socio-cultural influences of palm leaf manuscript collections in Bali in the development of document image analysis system. They affect its design di-

rectly from the beginning of the DIA pipeline. The first obstacle is the difficulty to collect the manuscript images as research corpus. In the next step, the need to construct the manuscript ground truth dataset will be directly affected by the limited number of Balinese philologists. In technical term, this condition will limit the performance evaluation of the future developed systems. For the development of DIA methods, the physical condition of the palm leaf materials and the complexity of the Balinese script become a perfect challenging combination for the researchers. The corpus of Balinese palm leaf manuscript will serve as a good dataset to test the robustness of the existing DIA methods. First, the image processing methods are not straightforward to be applied on this condition of manuscript images. Secondly, as one of the family of Brahmi script, the Balinese script is naturally complex which contains many glyph combinations and writing positions. The fact that the Balinese script is also a syllabic script makes the text transliteration more challenging. In the next chapter, the detailed description of the existing methods for each task from the state-of-the-art of DIA system will be presented. It starts from the binarization, the text line segmentation, the isolated glyph recognition, to the text transliteration.

Chapter 4

State-of-the-art of Document Image Analysis (DIA) System

This chapter concentrates on the presentation of all existing methods for each task from the state-of-the-art of DIA system. The global overview of DIA system and the ground truth construction are firstly given. The more detailed description of existing methods for each task in the DIA pipeline is then presented, starting from the binarization, the text line segmentation, the isolated glyph recognition, to the text transliteration.

4.1 Global Overview of DIA

During the last three decades, the field of DIA has become one of the most popular research areas and has made great progress in achieving the results and performances of its research methodologies. There are several things that support and drive this rapid development and progress. First, the existence of the main research object for the field of DIA, in this case it is the collection of the document itself, as described in the Chapter 1 and 2 of this dissertation. The existence of such a large collection of documents does not only bring precious values to the social sciences such as history, literature, and culture. The distribution and duplication of these documents through digitization has a wide effect in determining the DIA task definition. The effort to digitize the documents was only able to solve the problem of minimization of storage, keeping the contents from physical damage of documents, facilitating the distribution and dissemination of documents. DIA is needed to enable the extraction of all the explicit or implicit information contained in the document into a more structured representation [28]. It indirectly brought the main problems and challenges for research in the field of computer science which sparked the interest in DIA research. Secondly, the development and evolution of techniques and research methods in the field of DIA are triggered by the increasing complexity of the problems to be solved from a document. DIA seems to be a special field with its own key components [31]. This is where there is a merger between big and fundamental concepts of computer science such as image processing, machine learning and artificial intelligence to solve the problems. DIA is also supported by natural language processing and knowledge representation. It makes the field of DIA increasingly challenging and popular among researchers.

To be able to see and fully understand the position of problems and the evolution of DIA techniques and methods, the structure and classification of DIA can be seen from several points of view as it will be explained in the following sub sections.

4.1.1 Type of document

At the beginning of its emergence, DIA's research only dealt with very simple documents that only contained text, typed or machine printed, in Latin script and English. A text document contains paragraphs, text lines, word and characters. For text documents, the DIA's initial challenge was for typed or printed text with single fonts or multi fonts. It eventually switched to handwritten text or in more generic term, the unconstrained (non cursive or cursive) handwritten text.

As DIA research developed, it began to include documents containing graphs, such as tables, forms, logos, stamps, symbols or lettrines[1]. Some of the most complex documents with mixed text and graphic content such as maps, building plans, music scores and engineering drawings are now becoming the research objects for researchers in the field of DIA. The irregularities of combination of text and graphic contents in documents are increasingly becoming the latest challenge in DIA research when the object of research turns to a collection of ancient historical documents. Automatic recognition of historical documents is very challenging [29]. Techniques and methods for analyzing text documents and graphic documents can no longer be separated, but must be jointly analyzed

[1] An (ornamental) initial letter larger than the size of the text it accompanies (https://en.oxforddictionaries.com/definition/lettrine)

in a single document.

To observe more clearly the various types of documents that are becoming the object of DIA research, we can refer to various types of document datasets that have been published in the last few years, from the very popular and frequently used dataset like George Washington dataset from the IAM Historical Document Database (IAM-HistDB)[2] until some latest datasets which were proposed for DIA competition in recent years. For more complete and detailed DIA dataset examples, please refer to the TC-10[3] and TC-11[4] IAPR sites. We mention some examples of datasets as follows.

- **IAM datasets**[5]
  - IAM Handwriting Database
  - IAM On-Line Handwriting Database
  - IAM Online Document Database (IAMonDo-database)
  - IAM Historical Document Database (IAM-HistDB)
    - Saint Gall Database - 9th century, Latin
    - Parzival Database - 13th century, German
    - Washington Database - 18th century, English
- **Tobacco800 Complex Document Image Database**[6] [55, 56, 57]
- **Datasets from Prima Research**[7]
  - IMPACT Digitisation Centre of Competence Dataset
  - Layout Analysis Dataset A realistic contemporary document dataset.
  - Natural History Museum Lepidoptera Cards from the UK's Natural History Museum lepidoptera index.
  - Europeana Newspapers Project Dataset Newspapers from Europe's major libraries
  - Census 1961 Project Dataset Example images from the Census 1961 digitisation project
  - IMPACT Project Dataset Historical Books and Newspapers from Europe's major libraries
  - RDCL2017 ICDAR2017 Competition on Recognition of Documents with Complex Layouts - RDCL2017
  - REID2017 ICDAR2017 Competition on Recognition of Early Indian printed Documents - REID2017
  - RDCL2015 ICDAR2015 Competition on Recognition of Documents with Complex Layouts

[2] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database
[3] http://iapr-tc10.univ-lr.fr/index.php/resources/dataset-and-software
[4] http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List
[5] http://www.fki.inf.unibe.ch/databases
[6] http://www.umiacs.umd.edu/~zhugy/tobacco800.html
[7] http://www.primaresearch.org/datasets

  - HNLA2013 ICDAR2013 Competition on Historical Newspaper Layout Analysis
  - HBR2013 ICDAR2013 Competition on Historical Book Recognition
  - HDLAC2011 ICDAR2011 Historical Document Layout Analysis Competition

- **Marmot Dataset**[8]

  - Dataset for table recognition
  - Dataset for math formula recognition
  - Dataset for math formula identification in Chinese documents
  - Dataset for layout analysis of fixed layout documents

The recent datasets not only provide more complex documents, but they also offer more specific datasets which are intended for very specific tasks with special challenges in DIA. These kinds of datasets help a lot in benchmarking and evaluating some specific methods of DIA to increase the performance.

4.1.2 Level of processing

Because basically DIA research is conducted on image data, naturally the lowest level of the DIA process is done at the pixel level of the document image. This pixel level is the lowest entity level of a document image containing only gray level or color information of a single physical point on the document image. For some initial problems of DIA with types of document images that are still relatively simple, processing at this pixel level is enough to provide optimal performance. For example, separating the text and background of the document with bimodal gray level distribution or a fairly homogeneous color from each part of the text or background. In more complex problems, processing only at the pixel level is not sufficient. Therefore, higher level processing is proposed by using the primitive features as the descriptors of the document image. This feature level or descriptor serves to provide additional information about the contents of a document image by formulating some basic information from lower entity levels. A document image feature or descriptor is a more compact, solid and robust representation of some of the basic attributes and properties of a document image. For example, the connected component analysis is a preliminary feature descriptor processing, stroke analysis is a further level of processing in text document while line or curve and shape analysis is a further level of processing for graphic document [31, 30]. Other higher levels of processing for DIA for example structure level, document level and corpus level [31]. Some basic feature levels will be grouped to define the structure level of documents like words and text lines in text documents, or legends on graphic documents. The document level will be focused on the page layout analysis of the text document or the graphic interpretation for the graphic documents. Finally, the processing of each document in a collection will define the corpus level for searching, classifying, indexing and retrieval process of the documents. These different levels of processing will be reflected more clearly in different DIA applications in Sub Section 4.1.4.

[8] http://www.icst.pku.edu.cn/cpdp/data/marmot_data.htm

4.1.3 Data driven approach or model driven approach

In further development, the image features or descriptors in the data driven approach are completed by some semantic properties of the data. The DIA method is no longer based only on physical data, but it is now supported by a model. The symbolic representation on document images is not only to be extracted, but the implicit information in the document has to be interpreted. With a model driven approach, DIA now contains the document image understanding or document image interpretation concept.

4.1.4 Type of application and the processing pipeline

Optical Character Recognition (OCR) is actually just one of many tasks performed on a DIA system [28]. Although it can be said that OCR is the pioneer of the DIA field, there are now so many types of DIA tasks with successful applications. The solution of a problem in the DIA project does not often consists of a single processing task, but it consist of a sequence of processing tasks from several DIA tasks to be performed in a particular pipeline. Optimization of all tasks in this pipeline depends on the type of application to be built and also depends on the condition and characteristics of the document corpus as the object of interest. For each application, the DIA pipeline task should be customized. More detailed description about the pipeline of DIA will be given in Section 4.3.

4.1.5 Other DIA tasks for evaluation support and experimental protocol

The progress of research in the field of DIA is not only supported by many techniques and methods proposed to directly solve a problem found in a document image. However, in the DIA research community, there is also the need to work on the support for testing and evaluation of the techniques and methods developed. Some of DIA's tasks to complete the testing and evaluation protocol are as follows.

- **Document corpus collection and standard dataset construction.** More developed DIA research brings more specific and sharper DIA tasks and it is usually done for document collections that also have certain special characteristics. Many attempts to test the DIA methods on the collections of documents with this particular characteristic further incur the increasing number of document collection campaigns. The process of collecting the corpus of this document is generally under the framework of the task or project of digitizing documents. Some important things that must be considered in this stage is the quality assurance of the resulting document images, image capturing method and image format standardization for the entire corpus from the same document collection.

- **Dataset ground truthing process.** After the corpus of the document collection is collected, the next task is to make the ground truth dataset for the DIA task which will be applied to the document corpus. The ground truth dataset is a reference dataset that contains the correct information or the truth or the answer of the problem of the task. This ground truth dataset will be used as a comparison to measure how true (correct) is the result of the proposed DIA task and method applied on the same document corpus. The dataset ground truthing process is often a very time and energy consuming task on DIA research. It is because in many cases, the dataset groundtruthing process should be done manually by a human. Some DIA projects already proposed the semi automatic dataset ground truthing process. But it depends on the availability and the condition of the basic corpus of the document.

The more detailed overview on the ground truth construction process will be given in Section 4.2 of this dissertation.

- **Development of dataset ground truthing tools.** To facilitate and accelerate the dataset ground truthing process, especially to optimize the possibility of using semi automatic processes, several ground truthing tools were developed in several DIA projects. These ground truthing tools are expected to bridge the manual process that must be done by humans with some automatic processes that have been able to be done by one or several DIA methods that have been developed previously. Ground truthing tool are platform that provides a set of interactions between humans and the automatic sub modules that have been previously available. Additions, corrections or improvements to ground truth data information will be made by humans from previously automatically generated ground truth data information.
- **Development of benchmarking, evaluation and validation measures and tools.** In order to evaluate the performance of the DIA methods, the evaluation measures and tools have to be developed. The evaluation measure should be able to reflect the behavioral patterns of the method being tested against the overall case that may be found in the document collection. The resulting quantitative measure should represent well the range between the best case and the worst case that may be found from the output of the method being evaluated. In some DIA tasks, in addition to quantitative measures, visual observations are also needed to assess the performance of the tested method. For that reason, interfaces that can help humans in doing a visual observation of the results of DIA methods being performed are needed.

Some examples of ground truthing and benchmarking tool are Pink Panther as a tool to create segmentation ground truth files and benchmarking page segmentation algorithms [58], Aletheia[9] can be used to annotate textual content on the images [59], GEDI[10] as a groundtruthing environment for document images, and LabelMe[11] is open annotation tool to label objects on the images, PixLabeler offers an interface for pixel-level labelling of elements in document images [17].

4.2 Ground Truth Construction

Ground truth data construction is an important part in DIA researches. The quantitative measurements in evaluation protocol for a DIA method can only be done if the ground truth data are available. Depending on the availability and the condition of the basic corpus of document images, there are two options to create the ground truth data, by manual approach and semi-automatic approach. There is no fully automatic approach in constructing the ground truth as it will be the DIA method itself which should be evaluated.

4.2.1 Manual Approach

Document images are naturally a visual object. Most of the DIA tasks which are applied to the document images are intended to imitate the human vision capability in segment-

[9] http://www.primaresearch.org/tools
[10] http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=53
[11] http://labelme.csail.mit.edu/Release3.0/

ing, associating, recognizing and interpreting all parts or components on the document. It is why by definition, evaluating the performance of the DIA tasks means comparing the results to human vision result. The ground truth data which represent the truth from human vision should then be done manually by the human himself.

Constructing ground truth data by a fully manual approach is still reasonable when the number of document images to be ground truthed is still in small scale quantity and the time needed to construct the ground truth for one image is also reasonable. For example, for the creation of ground truth bleed-through document image database [60] of only 25 recto/verso image pairs, the ground truth foreground images were created manually, by drawing around the outline of foreground text on both recto and verso sides. These outline layers were then extracted from the images and filled in to create binary foreground text images. Another example, for binarization task, the manual ground truth binarized images were created by labeling the foreground and background parts of the image, manually pixel by pixel. In text recognition task, with the manual approach, the laborious works should be done first to manually trace the text line segments or maybe to a deeper level of word or character segments and then to annotate and to transcribe the segment images into their associated text format.

Ground truthing the documents manually is very costly in terms of man-hours [29, 51]. The manual ground truthing process can surely be faster by asking more people as ground truther, under condition that there exist those people who are eligible or expertized to do that works. Nowadays, for some DIA tasks which do not need certain specific level of expertise, the crowdsourcing method is very promising to be proposed in order to get a larger scale of ground truthed data in faster time.

Today, most of the ground truthing processes are not done totally manually anymore. At least it has been assisted by some tools to speed up the process, even though that tool does not give a decision about the truth of the data being ground truthed.

4.2.2 Semi-automatic Approach

The semi-automatic approach proposes a faster way to create the ground truth data by using some automatic tools or methods which provide the preliminary decision about the truth of the data being ground truthed. The automatic parts can be applied in the initial step or it can be inserted in the middle step of the ground truthing process. For example, for an OCR system with no or less available transcribed data, the framework of OCRoRACT [51] and anyOCR [29] proposed an approach to minimize the requirement of a language expert for manually transcribing documents. The semi-correct ground truth of the Unicode for character clusters are identified by the language expert after a semi-automatic text line and character segmentation. We decribe in this sub section the examples of semi-automatic approach to create the binarized ground truth images for the binarization task because the creation of binarized ground truth images can facilitate the creation of glyph or text line segmentation ground truth dataset.

Manual creation of the ground truth binarized images (e.g. with PixLabeler application [17]) is a time-consuming task. Therefore, several semi-automatic frameworks for the construction of ground truth binarized images have been presented [13, 12, 61, 62] to reduce the time of ground truthing process. The human intervention is required only for some necessary but limited tasks.

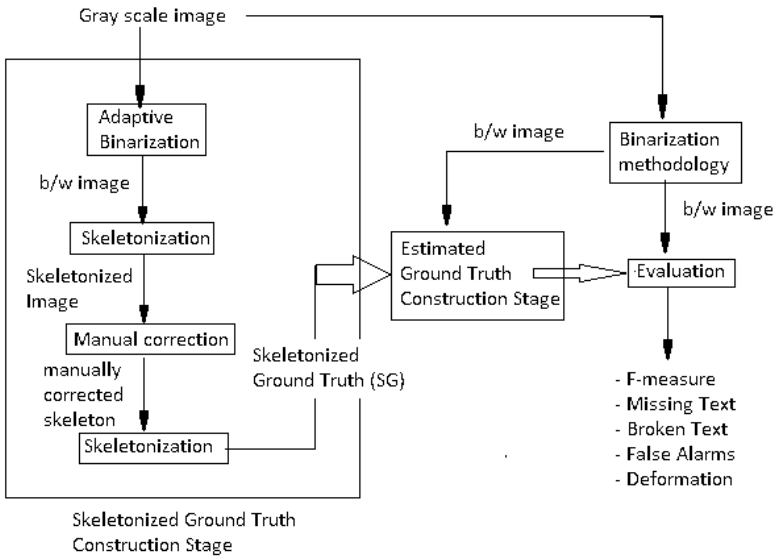


Figure 4.1: Frameworks For Construction of Ground Truth Binarized Images in [12]

The construction of ground truth binarized images was proposed in [12]. In this framework, the manual process takes part in the construction of a skeletonized ground truth image, and the automatic works are found in the initial step of binarization and in the final step of construction of estimated ground truth image. This framework consists of several steps: initial binarization process with an adaptive binarization technique, skeletonization of the characters to generate one pixel wide text (character), manual correction of skeleton to remove artifact and spurious parts and to complete the incomplete skeletonized character, and then second skeletonization after manual correction process is applied to guarantee that the ground truth skeletonized text are only one pixel wide. The estimated ground truth image is then fully constructed by dilating repetitively the corrected skeleton image, constrained by the character edges (detected using Canny algorithm [63]) and the binarized image under evaluation. The skeleton is dilated until half of the Canny edges intersect each binarized component. The detailed algorithm in pseudo code can be found in [12]. In this method, poor quality of initial binarized image will directly affect the result of the estimated ground truth. The ground truth image constructed strongly depends on the binarized image used as a constraint during the dilation process of the skeleton.

For the DIBCO competition series [2], the ground truth binarized images are constructed using a semi-automatic procedure described in [13]. The ground truth binarized images used for the DIBCO competition series are constructed with a modified procedure [13] as illustrated in Figure 4.2. In this procedure, the conditional dilation step of the skeleton is constrained only by Canny edge image, without any initial binarized image.

This procedure is adapted and improved by some other works on the construction of ground truth binarized images. For instance, in [3], a similar method is used to create ground truth of a large document database. In [61], in order to save user time in manual modification process by expert, two features of phase congruency are used to pre-process Persian heritage images to generate a rough initial binarized image. In [62], the ground truth binarized image of a machine-printed document is constructed by segmenting and clustering the characters during the foreground enhancement step. The user can manu-

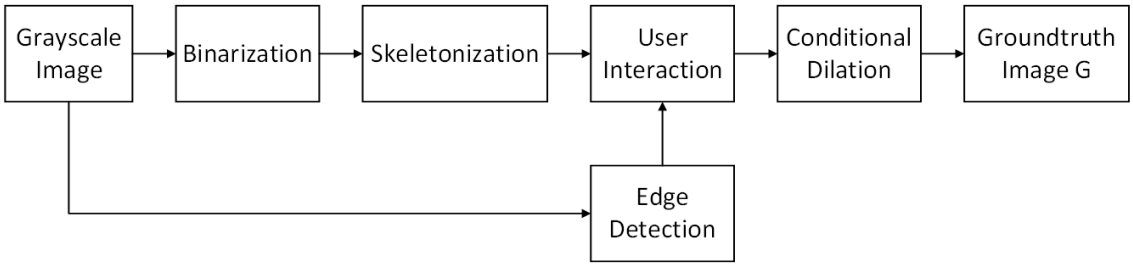


Figure 4.2: Ground truth construction procedure used for DIBCO series [13]

ally add and remove character model assignments to degraded character instances.

4.3 Methods and techniques for DIA pipeline

In this section, the methods for each DIA task for palm leaf manuscripts are presented, starting from the binarization, the text line segmentation, the isolated glyph recognition, to the text transliteration. For the binarization task, the methods from the state of the art to the latest methods in the binarization competition are described. Some text line segmentation methods for binarized and grayscale images are presented, including the seam carving method and the recent text line segmentation method for palm leaf manuscripts [64]. For the isolated character/glyph recognition task, the methods are reported from the handcrafted feature extraction method, the neural network with unsupervised learning feature to the Convolutional Neural Network (CNN) based method. And finally, the LSTM based method is described for the transliteration task of the palm leaf manuscripts.

4.3.1 Binarization

Binarization is widely applied as the first preprocessing step in image document analysis [5]. Binarization is a common starting point for the document image analysis pipeline, converting gray image values into binary representation for background and foreground, or in more specific definition, text and non-text, which is then fed into further document processing tasks such as text line segmentation and optical character recognition (OCR). The performance of the binarization techniques has a great impact and directly affects the performance of the recognition task [12]. Non optimal binarization methods produce unrecognizable characters with noise [21].

Based on the choice of the thresholding value, binarization methods can generally be divided into two types, global binarization and local adaptive binarization [21] (Figure 4.3). Some surveys and comparative studies of the performance of several binarization methods have been reported [12, 7]. A review of evaluation of optimal binarization techniques for character segmentation in historical manuscripts was presented in [22]. A binarization method that performs well for one document collection, may not necessarily be applied to another document collection with the same performance [5]. For this reason, there is always a need to perform a comprehensive evaluation of the existing binarization methods for a new document collection that has different characteristics, for example historical archive documents [7].

In this sub section, we describe several alternative binarization algorithms for palm leaf manuscripts. We compare some well-known standard binarization methods, and some binarization methods that are promising experimentally for historical archive doc-

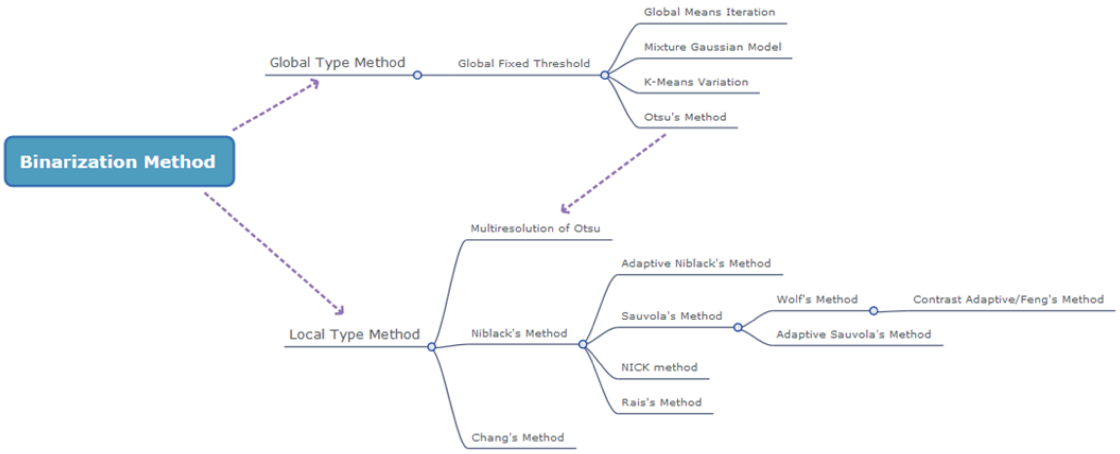


Figure 4.3: Types of Binarization Method

uments, not specifically for images of palm leaf manuscripts. We also describe the binarization methods from the binarization challenge of ICFHR competition[12] [23].

4.3.1.1 Global Thresholding Methods

Global thresholding is the simplest technique and is the most conventional approach for binarization [5, 6]. A single threshold value is calculated from global characteristics of the image. This value should be properly chosen based on a heuristic technique or a statistical measurement to be able to give a promising optimal binarization result [7]. It is widely known that using a global threshold to process a batch of archive images with different illumination and noise variation is not a proper choice. The variation between images on foreground and background colours on low quality document images gives unsatisfactory result. It is difficult to choose one fixed threshold value which is adaptable for all images [7, 8]. Otsu's method is a very popular global binarization technique [5, 6]. Conceptually, Otsu's method tries to find an optimum global threshold on an image by minimizing the weighted sum of variances of the objects and background pixels [5]. Otsu's method is implemented as standard binarization technique as a built-in Matlab function called graythresh[13].

4.3.1.2 Local Adaptive Methods

To overcome the weakness of the global binarization technique, many local adaptive binarization techniques were proposed, for example Niblack's method [5, 7, 6, 8, 4], Sauvola's method [5, 7, 6, 8, 4, 9], Wolf's method [8, 4, 65], NICK's method [4], and Rais method [5]. The threshold value in local adaptive binarization technique is calculated in a smaller local image area, region or window. Niblack's method proposed one of the well performed local thresholding computation based on the local mean and local standard deviation of a rectangular local window for each pixel on the image. The rectangular sliding local window will cover the neighborhood for each pixel. Using this concept, Niblack's method was reported to outperform many thresholding techniques and gave optimal acceptable results for many document collections. However, this method

[12] amadi.univ-lr.fr/ICFHR2016_Contest
[13] https://fr.mathworks.com/help/images/ref/graythresh.html

presents a drawback. Niblack's method works well only on the text region, but it is not well suited for a large non-text regions on the image. The absence of text on the local areas led Niblack's method to detect noise as text. The suitable window size should be properly chosen based on the character and stroke size which may vary on each image. Many other local adaptive binarization techniques were then proposed to improve the performance of the basic Niblack's method. For example, Sauvola's method is a modified version of Niblack's method. Sauvola's method proposes a local binarization technique to deal with light texture, big variations and uneven illumination. The improvement from Niblack's method is on the use of adaptive contribution of standard deviation in determining local threshold on the gray values of text and non-text pixels. Sauvola's method processes the image in NxN adjacent and non-overlapping blocks separately. Wolf's method tried to overcome the problem of Sauvola's method when the gray values of text and non-text pixels are close to each other by normalizing the contrast and the mean gray value of the image to compute the local threshold. However, a sharp change in background gray values across the image decreases the performance of Wolf's method. Two other improvements of Niblack's method are Nick's method and Rais's method. Nick's method proposed a threshold computation derived from the basic Niblack's method while Rais's method proposed an optimal size of window for the local binarization.

4.3.1.3 Methods from ICFHR 2016 Competition Challenge 1: Binarization of Palm Leaf Manuscript Images

This competition provides an opportunity for a variety of proposed binarization methods that may be applied to Balinese palm leaf manuscripts. The results of this competition can provide a broader basis for comparative studies for the development of the binarization method for Balinese palm leaf manuscripts. The top two proposed methods in Binarization Challenge for ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts are the training based binarization methods [23]. The best method in this competition (ICFHR G2) employs a Fully Convolutional Network (FCN). The second best method (ICFHR G1) uses two neural network classifiers. The details of all the submitted methods are described as follow.

Method from Group 1 (ICFHR G1): First, they use two neural network classifiers C1 and C2 to classify each pixel whether it is background or not. Two binarized images B1 and B2 are obtained in this step. C1 is a rough classifier which tries to detect all the foreground pixels while probably making mistakes for some background pixel. C2 is an accurate classifier which should not classify the background pixel as foreground pixel while probably missing some foreground pixels. These two binary images are then joined to get the final classification result. In this step, they use each foreground pixel in C2 as a seed and find all the foreground pixels in B1 and B2 which is connected with the seed pixel. All the foreground pixels in B1 which are not connected to any seed pixel in B2 are eliminated. The output image in this step is denoted as B3. Finally, the skeleton image of B3 is extracted and the dilated skeleton image is the final binary image of the input image. In the training stage, they set the structure of C1 and C2 as 124x100x100x1 and 221x100x100x1, respectively. All the transfer functions were set as the sigmoid function. In C1, the input feature f_i of each pixel P_i is the combination of the RGB values of P_i and the subtractions between the gray value of P_i and the neighborhoods P_j, j=1,...,121. Neighborhoods are the pixels in the 11x11 window with P_i as the window center. In C2, the feature f_i of each pixel P_i is the combination of the gradient features [66] around P_i and the subtractions between the gray value of P_i and the neighborhoods P_j, j=1,...,121.

In the experiment, they used a 45x45 image block around P_i to extract the gradient feature. The image block is divided into 5x5 sub-blocks and 4 direction gradient features are extracted in each sub-block.

Method from Group 2 (ICFHR G2): They employ a Fully Convolutional Network (FCN). It takes a color sub image as input and generates the probability that each pixel in the sub image is part of the foreground as output. The FCN is pre-trained on normal handwritten document images with automatically generated "ground truth" binarizations (using the method of Wolf et al [65]). The FCN is then fine-tuned using DIBCO and HDIBCO competition images and their corresponding ground truth binarizations. Finally, the FCN is fine-tuned again on the provided Balinese Palm Leaf images. For inference, the pixel probabilities of foreground are efficiently predicted for the whole image at once and thresholded at 0.5 to create a binarized output image.

Method from Group 3 (ICFHR G3): The submitted method adopts Lu et al.'s method [67] to estimate the document background based on the polynomial smoothing. The background is estimated by fitting a polynomial function vertically and horizontally. The estimated background is then used to compensate the contrast of the input document image as described in [67]. Subsequently, they apply the local contrast evaluated by the local maximum and minimum [68] to further suppress the background variation on the normalized image. The text stroke edge pixels in the processed image can then be segmented by a global threshold. Finally, the text pixels are classified based on the following criteria: 1) There are text stroke edge pixels existing within a local window of the text pixel, 2) The intensity of the text pixel should be similar to the text stroke edge pixel.

Method from Group 4 (ICFHR G4): They proposed a method named "Hue segmented local contrast binarization". The palm Leaf RGB image is converted into a HSV image. Hue is used to segment the foreground and the background from the image. The boundary between the foreground and the background is used as a boundary image. The RGB image is combined together to form a gray image using the peak position in the individual histogram as the scaling parameter. The gray image is converted to a local contrast image using a Weiner filter. The local contrast image is binarized using Sauvola's threshold N=32, k=0.1, and R=128. The binarized image is filtered using the boundary image and the threshold based on the area, the height, and the width of individual connected component in the image.

4.3.1.4 Conclusions

As a common starting point for DIA pipeline, the binarization has a great impact and directly affects the performance of the recognition task. The conventional approaches of binarization use the threshold value to separate foreground and background. The threshold value can be calculated globally from the whole image. This technique is simple, but it gives unsatisfactory result in low quality images with variation in colors. To deal with the noise variation and uneven illumination, the local adaptive binarization techniques are proposed. With these methods, the threshold value is calculated locally in each smaller local image area, region or window. Even though these methods overcome the problems of image variation, they present a drawback for the image with large part of non text area. In most cases, noise is detected as text. The training based binarization methods require binarized image samples to be trained with classifiers. The recent binarization method uses neural network as classifiers or employ convolutional networks.

4.3.2 Text Line Segmentation

An important entity in a document image is a text line. A text line is normally composed by some words which are arranged in such spatial position, so it represents the reading order of all words of the document in the horizontal direction. Vertical position of some text lines also provides an important information about a paragraph which represents the layout of the document. There are cases where the text lines are vertical, for example in Japanese. Similar reasoning can be applied for vertically oriented texts. Segmentation of a document image into physical spatial entities such as text lines, words, and characters is often performed prior to recognition step of an OCR system [33, 53, 54, 52, 69, 70, 71]. The segmentation based text recognition method requires prior segmentation processing of the document image into text line segments, word segments, or character segments.

Many methods of text line segmentation for handwritten document image have been already proposed [52, 72, 73, 74, 75, 76]. Some works deal directly with the text line and character segmentation and recognition [53, 69, 70]. A survey of text line segmentation methods for historical documents is given in [77]. But most of those methods basically still depend on the binary image of the document. Some methods for text line or character segmentation directly applied from grayscale image have already been proposed [70, 71, 78, 79]. Some other methods used the combined information from both binary and grayscale image [53, 69]. In this dissertation, we investigated the performance of six promising text line segmentation methods by conducting the comparative experimental studies on the collection of palm leaf manuscript images. We describe six promising text line segmentation methods that are used in our experimental studies. Three methods work on binary images: the Adaptive Partial Projection (APP) line segmentation approach, the A* Path Planning approach, the shredding method, and our proposed energy function for shredding method. Three other methods that can be directly applied on grayscale images are also investigated: the Adaptive Local Connectivity Map (ALCM), the seam carving based method and the Adaptive Path Finding Method.

4.3.2.1 Methods for Binarized Images

4.3.2.1.1 Adaptive Partial Projection (APP)

The APP line segmentation approach was proposed by Chamchong and Fung [80]. It is an improved technique from their previous work [49] by adapting the modified partial projection and smooth the histogram with recursion. The technique first constructs the global horizontal projection of the text image to determine the number and average positions of text lines and the average distance between two adjacent lines. These details will be used throughout as reference values. The whole image is then divided into vertical columns. The column size is estimated to be "3*average_char_width" as it is normally the size of a word. The average character width and height are automatically calculated from analysis on connected component of each binary image of the manuscript. The smoothed horizontal projection profile is extracted from each column, and the valleys of the profile are considered to be the base lines of that column. For each column, incorrect base lines are removed, and new base lines are inserted based on the referenced values mentioned above. The approach also deals with connected components that spread over multiple lines by recursively dividing the column, in which those components belong, into two, and by traversing up and down from the old base lines until it reaches a more appropriate position. The base lines of all columns are joined together to form separating lines (Figure 4.4).

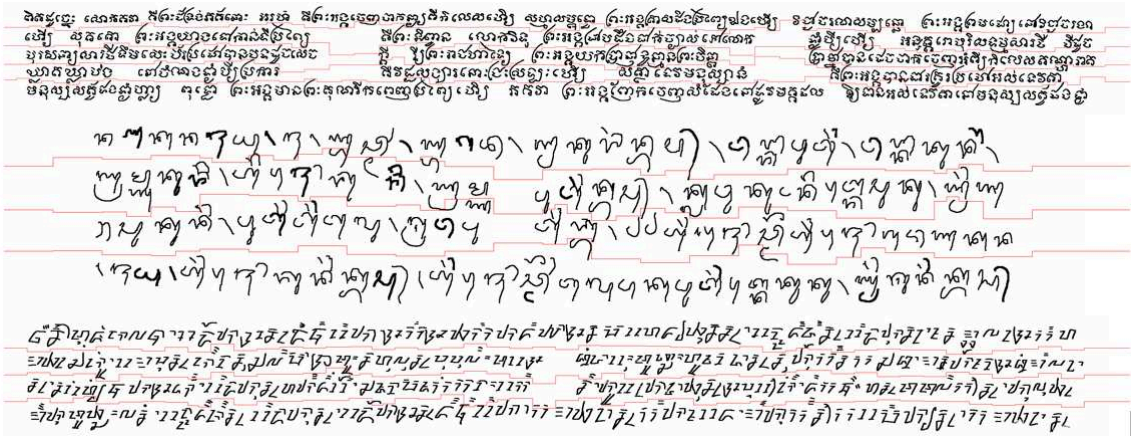


Figure 4.4: From top to bottom: Some results of the APP approach on Khmer, Balinese, and Sundanese manuscripts

In our experiments, the method first calculates the number of lines based on the number of peaks of the global horizontal projection profile. To find the number of peaks, we have to smooth the projection profile using a moving average filter to remove spurious peaks. The window size of the filter depends on the average height of all connected components in the page. For Khmer and Balinese scripts, we used "avg_char_height/2" as the window size. However, for Sundanese scripts, since they contain lots of smalls connected components, we have to increase the window size to "avg_char_height" to obtain better results.

4.3.2.1.2 Shredding Method

The shredding method was proposed by Nicolaou and all [52]. This technique tries to shred the image into text lines from one side of the image to the other side by following the white-most and black-most paths. This approach consider a topological assumption that for each text line, there exists a path from one side of the image to the other that traverses only one text line. The shredding method is applied on the binary image.

In the pre-processing stage, the binary image is blurred with a blurring filter that is based on the size of the estimated letter height from all connected component heights on binary image of the manuscript. The most frequent letter height found in all connected component heights on a binary image of the manuscript is used. Let I be the binary image, LH be the estimated letter height. The width of the blurring window is defined as BW=LH*8, and the height of the blurring window is defined as BH=LH*0.8. The blurring image B is defined as:

$$B(x,y) = \sum_{i=-BW/2}^{i=BW/2} \sum_{k=-BH/2}^{k=BH/2} I(x+i, y+k). \tag{4.1}$$

The size of BW and BH are defined in such a way that this operation blurs out the intra-characters and intra-words spaces while keeping the spaces between text lines (Figure 4.5). A recursive tracer function Tr is finally applied on the blurred image. The function Tr is defined as:

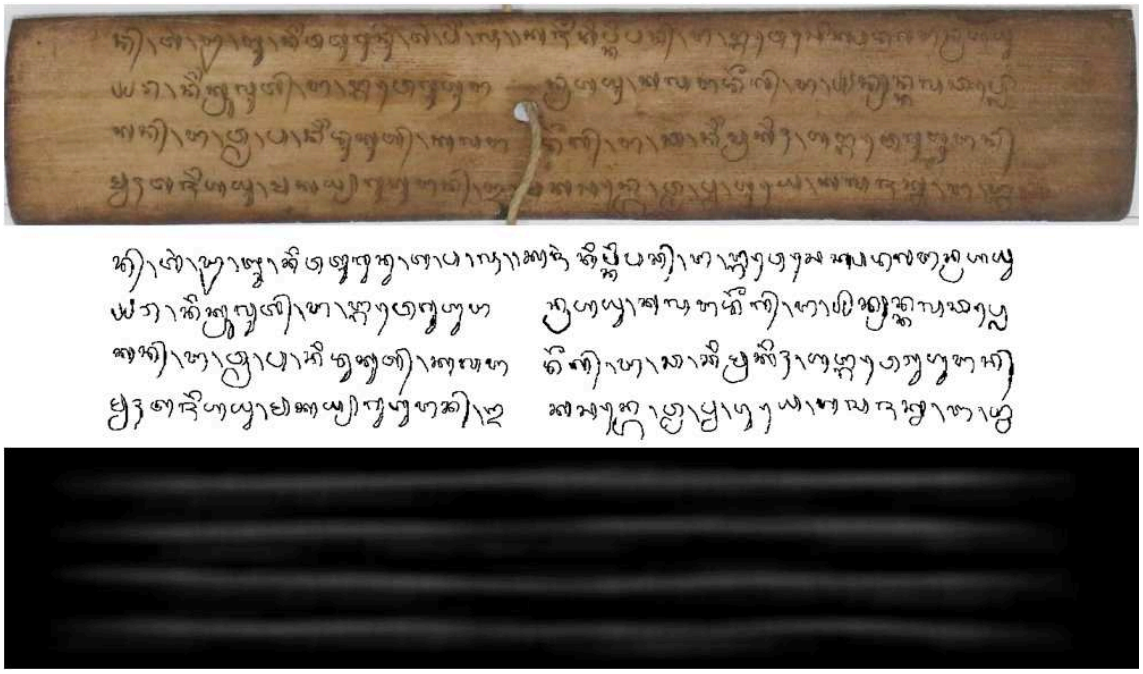


Figure 4.5: From top to bottom: original image, binary ground truth image, and blurred image of shredding technique

$$Tr_{k,B}(1) = k$$
$$Tr_{k,B}(n+1) = \begin{cases} Tr_{k,B}(n) - 1 \Rightarrow if : B(n, Tr_{k,B}(n) + BH/2) > B(n, Tr_{k,B}(n) - BH/2) \\ Tr_{k,B}(n) \Rightarrow if : B(n, Tr_{k,B}(n) + BH/2) = B(n, Tr_{k,B}(n) - BH/2) \\ Tr_{k,B}(n) + 1 \Rightarrow if : B(n, Tr_{k,B}(n) + BH/2) < B(n, Tr_{k,B}(n) - BH/2) \end{cases} \tag{4.2}$$

The function generates the shredded text line areas (Figure 4.6). The text line areas which are smaller than LH2 were filtered out. To detect the medial axis of text lines (Figure 4.7), the same recursive tracer function is applied on the inverted blurred image –B(x,y). The next step is assigning each connected component from the binary image input based on the intersection with line areas and line centers or medial axis.

4.3.2.1.3 A* Path Planning Approach

The A*PP line segmentation approach has been proposed by Surinta et al [81]. The objective of path planning is to compute the shortest path from a starting point to its destination avoiding obstacles along the way. A* (called A star) is one of the path planning algorithms which minimizes the travel costs between states from the starting state *s1* until the goal state *sn*. To solve the line segmentation problem, paths separating text lines need to be traced from the left side (starting state) to right side (goal state) of the text, and the foreground (black) pixels are viewed as obstacles. Due to some handwritten text components from adjacent lines being touching or overlapping, the goal state can be unreachable. A modified A* path-planning technique is proposed to allow the path to pass through such components. The method now works as follow. The position of the starting state and the goal state of each path is computed from the valley points of the smoothed y-projection profile histogram. Five cost functions are then combined to determine the traveling cost between states:

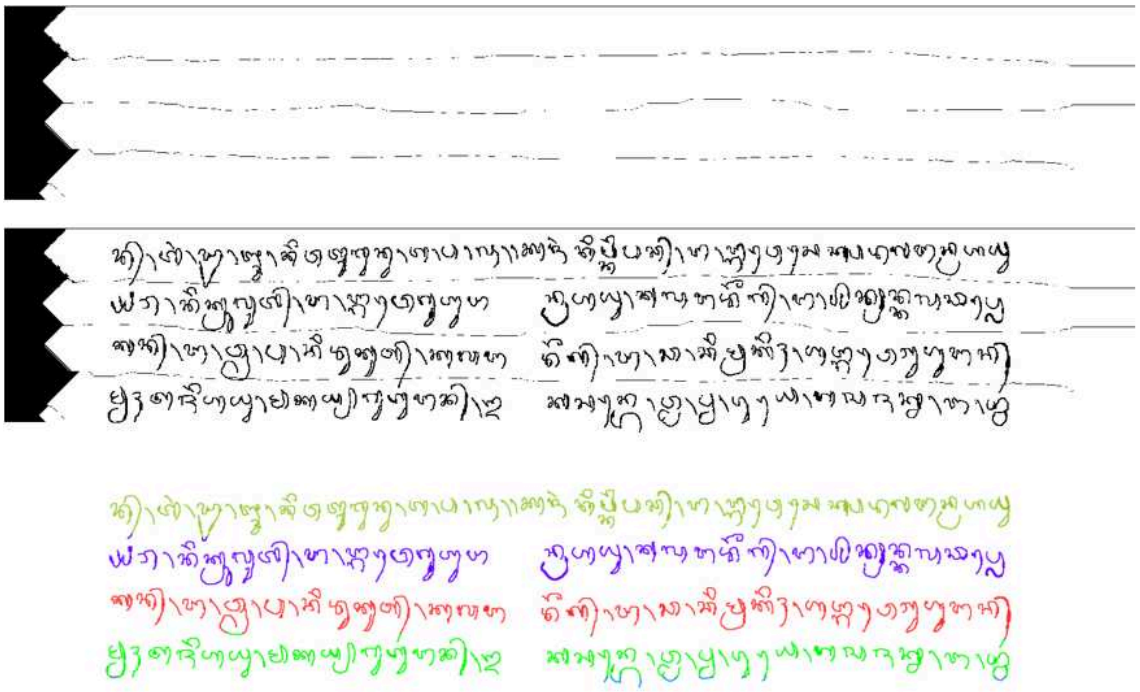


Figure 4.6: The shredded text line areas and the text line segments

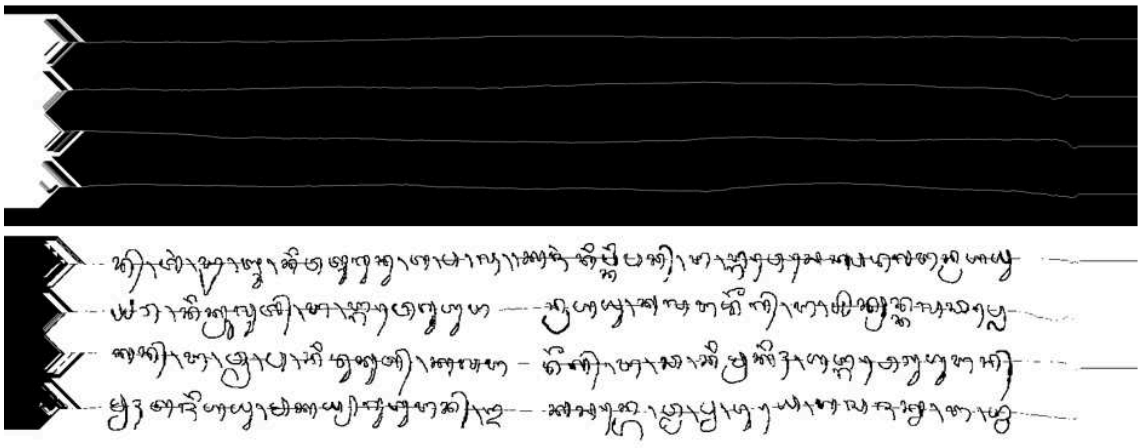


Figure 4.7: The detected medial axis of text lines

- Foreground distance cost functions D(n) and $D(n)^2$: control the path to stay along the gaps between foreground pixels. The two functions are defined as:

$$D(n) = \frac{1}{1 + min(d(n, n_{y_u}), d(n, n_{y_d}))} \tag{4.3}$$

$$D(n)^2 = \frac{1}{1 + min(d(n, n_{y_u}), d(n, n_{y_d}))^2} \tag{4.4}$$

  where $d(n,n_{y_u})$ and $d(n,n_{y_d})$ are the distances between the state n and the closest foreground pixel in the upward and downward direction respectively.

- Map-obstacle cost function M(n): gives penalty if the path has to pass through foreground pixels. M(n) returns 1 if the state n coincides with a foreground pixel, and it returns 0 otherwise.

- Vertical cost function V(n): prevents the path from deviating from the y-position of the starting state and the goal state. The function is defined as:

$$V(n) = \left| n_y - n_y^{start} \right|. \tag{4.5}$$

- Neighbor cost function N(si,sj): computes the shortest path between the starting state and the goal state. Like in the standard algorithm, this function returns 14 for diagonal directions (1.4 unit in diagonal direction, multiplied by 10 to round up its value), and it returns 10 (1 unit multiplied by 10) for other directions in the 8-directional movements.

The combined cost function $C(s_i,s_j)$ is defined as:

$$C(s_i, s_j) = c_d D(s_i) + c_{d2} D(s_i)^2 + c_m M(s_i) + c_v V(s_i) + c_n N(s_i, s_j). \tag{4.6}$$

The parameters c_d, c_{d2}, c_m, c_v, and c_n are tuned empirically.

There are two major drawbacks in this approach. The first downside is that the method assumes the y positions of the starting state and the goal state to be the same. Therefore, it does not work well with documents containing curved or slanted text lines. Another difficulty is caused by the tuning of the coefficient parameters in order to find the most efficient values to compute the final traveling cost $C(s_i,s_j)$.

In our experiment, the position of the starting state and the goal state are calculated separately in order to adapt to the skewness of the text lines. The y-positions of the starting states and the goal states are extracted from the y-projection histogram of the first one-third of the document and the last one-third of the document respectively. The vertical cost function now becomes:

$$V(n) = \left| n_y - \left(\left(1 - \frac{n_x}{l}\right) n_y^{start} + \frac{n_x}{l} n_y^{goal} \right) \right| \tag{4.7}$$

where $l=n_x^{goal}-n_x^{start}$ and (n_x,n_y), $(n_x^{start},n_y^{start})$, (n_x^{goal},n_y^{goal}) are the coordinates of the current state n, the starting state, and the goal state respectively. Normally, n_x^{start} is 0 and n_x^{goal} correspond to the end of the document page, so l is equal to the width of the document. The cost V(n) is now the vertical distance from the current state at position (n_x,n_y)

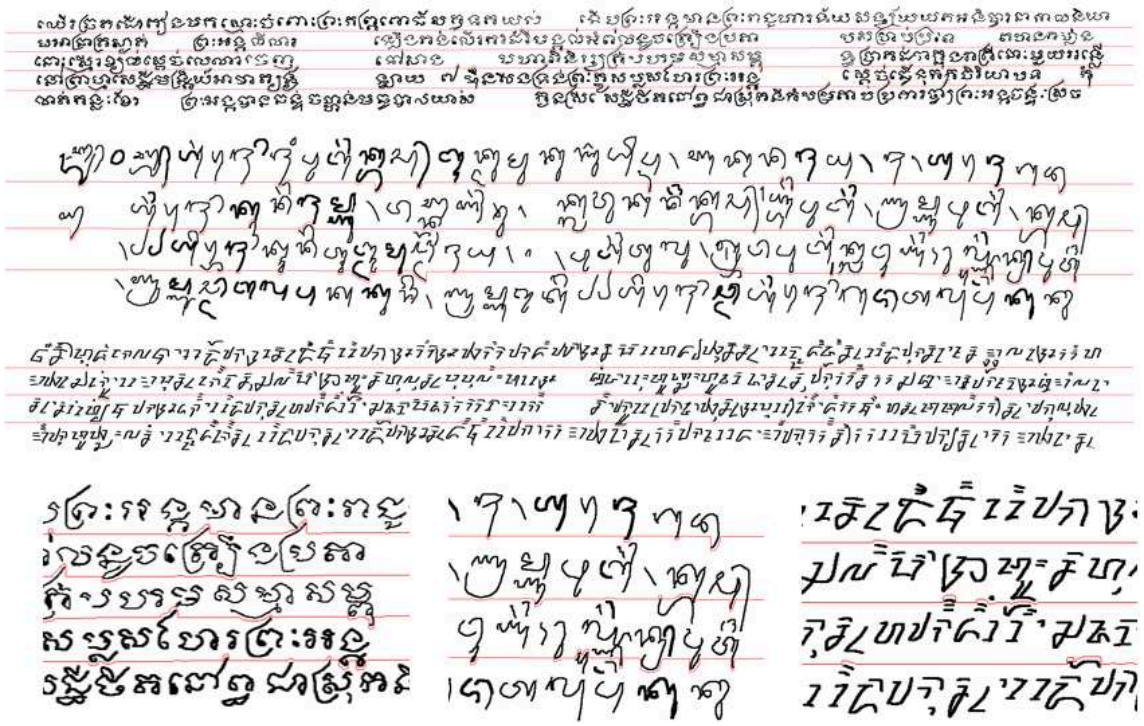


Figure 4.8: Example of results of A* Path Planning Approach

to the slanted line constructed from the two points (n_x^{start},n_y^{start}) and (n_x^{goal},n_y^{goal}). To improve the execution time, only five directional steps (S, SE, E, NE, E) are computed for neighbor cost function $N(s_i,s_j)$. The values of the parameters used on all data sets are: c_d=150, c_{d2}=0, c_m=50, c_v=5, and c_n=1 (Figure 4.8).

4.3.2.2 Methods for Grayscale Images

In this sub section, we investigate three line segmentation methods that are independent of binarization task. These approaches work directly on grayscale images.

4.3.2.2.1 Adaptive Local Connectivity Map (ALCM)

The ALCM method was proposed by Zhixin and all [71]. This method is considered as a transform based method and can be applied directly on grayscale images. It consists of the following steps. First, an ALCM map is generated from grayscale document using this following transform [71].

$$\begin{gathered} ALCM : f \rightarrow A \\ A(x,y) = \int_R f(x,y) G_c(t-x,y)dt \\ where \\ G_c(x,y) = \begin{cases} 1 \Rightarrow if : |x| < c \\ 0 \Rightarrow otherwise. \end{cases} \end{gathered} \tag{4.8}$$

The ALCM transform computes the cumulative intensity by adding up all the intensity values in a certain size of neighborhood of each pixels. The ALCM transformed image is a grayscale image that gives the preliminary information about the possible locations of text lines. The second step is the binarization of ALCM transformed image. It was

considered that the binarization of ALCM transformed image is easier than the original grayscale image because it consists of a clear bi-modal pixel distribution. The ALCM transformed image has more tolerence towards different binarization algorithms.

After the binarization of ALCM transformed image, a procedure to filter out the small pieces area is performed. It is based on the statistical size from all area in the binary image of ALCM. The full area of a text line is then generated by filling the preliminary text line area based on their upper and lower profile points. The binarization is finally performed by locally focusing only on the area of text lines. The last step is the connected component mapping and text line component collection. All connected components found in the binarized version of the document are mapped based on the locations of the text line area to make up the text line segments.

In our experimental studies, we computed the cumulative intensity in a neighborhood of size 2c, where c=100, as it was suggested to approximate c by the value of three times the average height of text. The scanning process to add up all intensity values was done twice, from left to right and right to left, and all cumulative intensities were finally re-scaled to range from 0 to 255 to produce a grayscale ALCM transformed image. For the next step of binarization of ALCM transformed image, we found that it is still difficult to binarize the ALCM transformed image from our manuscripts. The local adaptive binarization method of Sauvola [9] is applied with default values of k=0.1, R=128, and block_size=50x50. In most of the cases, the ascender and descender parts of the character make it difficult to separate two consecutive text line areas. Consequently, the preliminary position of text lines in a document is still hardly extracted (Figure 4.9). We finally filtered out the small areas whose height is less than 20 pixels (the half of the estimated text height) or with a width of less than 100 pixels (a half of the estimated word width). To make up the text line segments, we mapped up all connected components found in the binary ground truth image of the document, based on the intersection with the text line areas. If a connected component intersects more than one text line area, it will be assigned to the text line with the most intersection area. We did not perform any further post processing task in order to investigate the performance of the ALCM transform.

4.3.2.2.2 Seam Carving Based Method

The seam carving based method determines the segmentation path based on a defined seam map which is generated from a given energy function. Some schemes for text line segmentation based on the seam carving method have already been proposed [74, 82, 83, 84, 85]. By using the same basic idea for the seam carving method, the different schemes differ only on the choice of the energy function and in their proposed pre-processing and post processing steps.

Saabni and El-Sana proposed a language-independent text lines extraction using seam carving [83]. Their method is applied to the binary image. In the pre-processing step, they calculated the average height of connected component and classify them (according to their height) into four categories: additional strokes, ordinary average components, large connected components, and vertically touching components. In this method, the seam carving concept is applied to find the medial axis of the text lines. The Signed Distance Transform (SDT) is used as energy function to compute the energy map. The dynamic programming is finally used to compute the minimal cost seam that passes from the left side of the image to the right side. The collection of the connected components

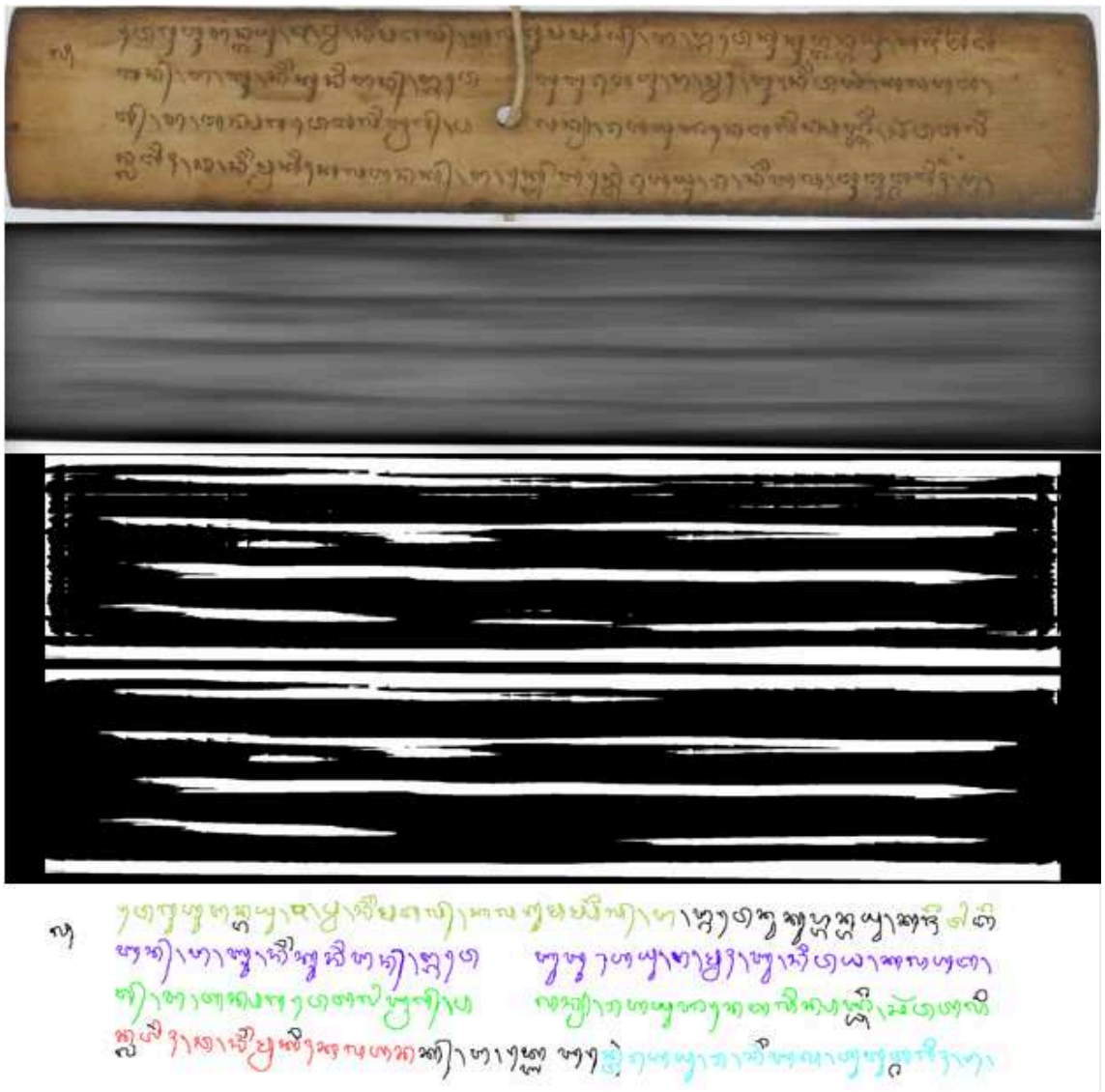


Figure 4.9: From top to bottom: original image, ALCM transformed image, the binarized image of ALCM tranformed image, the final binarized image of ALCM tranformed image after filtering process of small areas, and the text line segmentation

is then performed by applying some rules based on the position of the extracted medial axis of the text lines.

Some new approaches for text line segmentation based on the seam carving method that work directly on grayscale document images have been also proposed [82, 84]. In this method, two types of seams are calculated: the medial seams and separating seams. Christopher et al. [85] used the simple gradient magnitude function as the energy function. Arvanitopoulos and Süsstrunk [82] proposed a binarization free method based on a two-stage process: medial seam and separating seam computation. The approach computes medial seams by splitting the input page image into columns whose smoothed projection profiles are then calculated. The positions of the medial seams are obtained based on the local maxima locations of the profiles. The goal of the second stage of the approach is to compute separating seams with the application on the energy map within the area restricted by the medial seams of two neighboring lines found in the previous stage. The technique carves paths that traverse the image from left to right cumulating energy. The path with minimum cumulative energy is then chosen. Nikolaos and all [82] used the derivative image of the grayscale manuscript as the energy function, as follow:

$$E_{i,j} = \left| \frac{I^{\sigma}_{i,j+1} - I^{\sigma}_{i,j-1}}{2} \right| + \left| \frac{I^{\sigma}_{i+1,j} - I^{\sigma}_{i-1,j}}{2} \right| \tag{4.9}$$

where I^{σ} is the smoothed grayscale image with Gaussian filter of standard deviation. For both works, the seam map is generated by the function,

$$\begin{aligned} M(i,1) &= E(i,1) \\ M(i,j) &= E(i,j) + min \begin{cases} M(i-1,j-1) \\ M(i,j-1) \\ M(i+1,j-1). \end{cases} \end{aligned} \tag{4.10}$$

In the work of Abedelkadir et al. [84], the distance transform is adopted as the energy function to generate an energy map. In their case, the medial seams are determined by local minimum points of the seam map, and the separating seams is determined by maximum points. In order to generate an accurate energy map and produce robust seams, they used different weights for the horizontal and diagonal distances. The seam map is generated by the function,

$$M(i,j) = 2 * E(i,j) + min \begin{cases} \frac{1}{\sqrt{2}} * M(i-1,j-1) \\ 1 * M(i-1,j-1) \\ \frac{1}{\sqrt{2}} * M(i+1,j-1). \end{cases} \tag{4.11}$$

As the starting point in our experimental studies, we investigated the performance of basic seam carving scheme by minimizing the pre-processing and post-processing steps. We have tested the seam carving method on binary image and grayscale image. We have followed the generic basic scheme for the seam carving method which consists of three steps: calculate the energy map of the image by using an energy function, generate the seam map, and trace the minimal path by following the minimal cost provided by the seam map.

The distance transform is used as the energy function. The distance transform calculates the minimum distance of each pixel in the image from the nearest background pixel. In

Figure 4.10: From top to bottom: original image, energy map, minimum path, and text line segmentation result

our experiments, to calculate the distance transform for binary images, we used the Matlab function *bwdist* with the Euclidean distance metric, and for grayscale images, we used the Matlab function *graydist* with quasi Euclidean distance metric. The seam map is generated by using different weights for the horizontal and diagonal distances, and by using two passes from left to right and from right to left as proposed by Abdelkadir et al. [84]. We then calculated the final seam map from the average of two seam maps from two directions. We generated the separating seam paths by tracing the minimum seam cost from both directions. All separating seam paths define directly the text line areas (Figure 4.10). We did not perform any further post processing task.

Secondly, for comparison purpose against a more complete seam carving scheme, we also evaluated the seam carving method which was implemented by Nikolaos et a.l [82]. In their implementation, the medial seams are first computed based on the projection profile matching approach. The image is divided into some slices of column, and the So-

Figure 4.11: The medial seams and the separating seams of the manuscript of Figure 4.10

bel operator is applied to compute the edges of the image. For each slice of the image, the smoothed horizontal projection profiles are calculated independently. The local maxima locations of the profile in all slices are then detected and connected to create piece-wise linear seams that approximate the medial axis of the text lines in the manuscript page. The separating seams are then computed with these medial axes as constraint. The medial axes enforce the separating seam to pass between two consecutive text lines. The energy map is calculated from the derivative image of grayscale image smoothed with a Gaussian filter (Figure 4.11).

4.3.2.2.3 Energy Function for Shredding Method

After analyzing the jumping and the joining path problem on the seam carving method, we have proposed an energy function that can still represent a high energy value in an empty area within one text line. For each pixel, we define the energy value as the number of text (foreground) pixels in an ellipse area which is centered in that pixel. The size of the minor axis and the major axis of the ellipse area are based on the approximated average character height and width that are calculated from the connected component analysis. The minor axis of the ellipse area is set with the character height to limit the transfer energy between two text lines, in order to avoid the jumping and the joining path. The major axis of the ellipse area should be set large enough (n times of character width), to be able to transfer the energy from one side to the other side of a text line in an empty text line area.

Let I be a binary image of size (nb_row x nb_col) where a foreground pixel is 1 and a background pixel is 0, r_{major} and r_{minor} be the size major axis and minor axis of an ellipse area. The energy value of a pixel in row x and column y is defined as:

$$E(x,y) = \sum_{i=1}^{nb\_row} \sum_{j=1}^{nb\_col} I(i,j) * Ellipse(i,j,r_{minor},r_{major},x,y) \tag{4.12}$$

where Ellipse(i,j,r_{minor},r_{major},x,y) is a binary function to check whether the pixel I(i,j) is inside the area of an ellipse with minor axis r_{minor}, major axis r_{major}, and is centered on the coordinate (x,y), defined as:

$$Ellipse(i,j,r_{minor},r_{major},x,y) = \begin{cases} 1, if \left(\frac{(abs(i-x))^2}{r_{minor}^2} + \frac{(abs(j-y))^2}{r_{major}^2} \right) \leq 1 \\ 0, if not. \end{cases} \tag{4.13}$$

After generating the image of the energy function, we apply the recursive tracing function of the shredding method directly (see Sub Section 4.3.2.1.2). Figure 4.12 shows that the ellipse energy function can transfer the energy to fill the empty text line areas, and

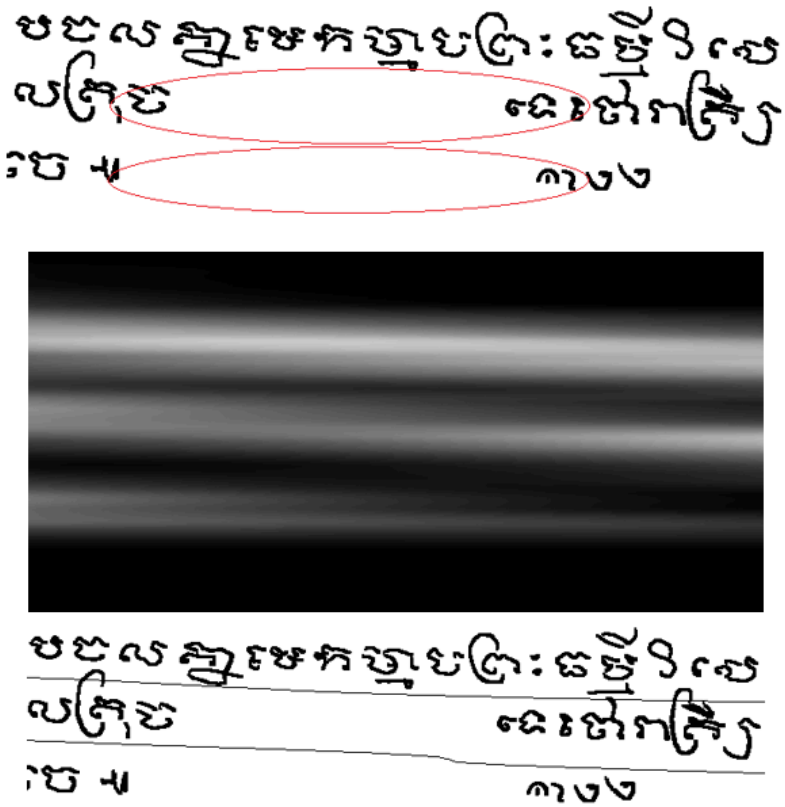


Figure 4.12: From top to bottom: Ellipse area in the empty text line area, image of the energy function, and the text line segmentation path generated from shredding method

then the recursive tracing function of the shredding method can easily separate the text line areas. We did not perform any further post processing task in order to investigate the performance of our proposed energy function for shredding method.

In our experiments, we used r_{minor} = *average character height* for all manuscript collection, r_{major} = 5 x *average character width* for the manuscript from Bali and Sunda, and for the Khmer manuscript r_{major} = 8 x *average character width* (Figure 4.13). These values have been defined empirically based on the existance of the empty text line area width in the manuscript collection. Khmer manuscripts usually have a wider empty text line area than the manuscripts from Bali and Sunda.

4.3.2.2.4 Adaptive Path Finding Method

This approach was proposed by Valy et al. [64]. The method takes a grayscale image of a document page as input. Connected components are extracted from the input image using the stroke width information by applying the stroke width transform (SWT) on the Canny edge map. The set of extracted components (filtered to remove components which come from noise and artifacts) is used to create a stroke map. Using column wise projection profiles on the output map, the estimated number and medial positions of the text line can be defined. To adapt better to skew and fluctuation, an unsupervised learning called competitive learning is applied on the set of connected components found previously. Finally, a path finding technique is applied in order to create seam borders between adjacent lines by using a combination of two cost functions: one penalizing the path that goes through foreground text (intensity difference cost function D) and another

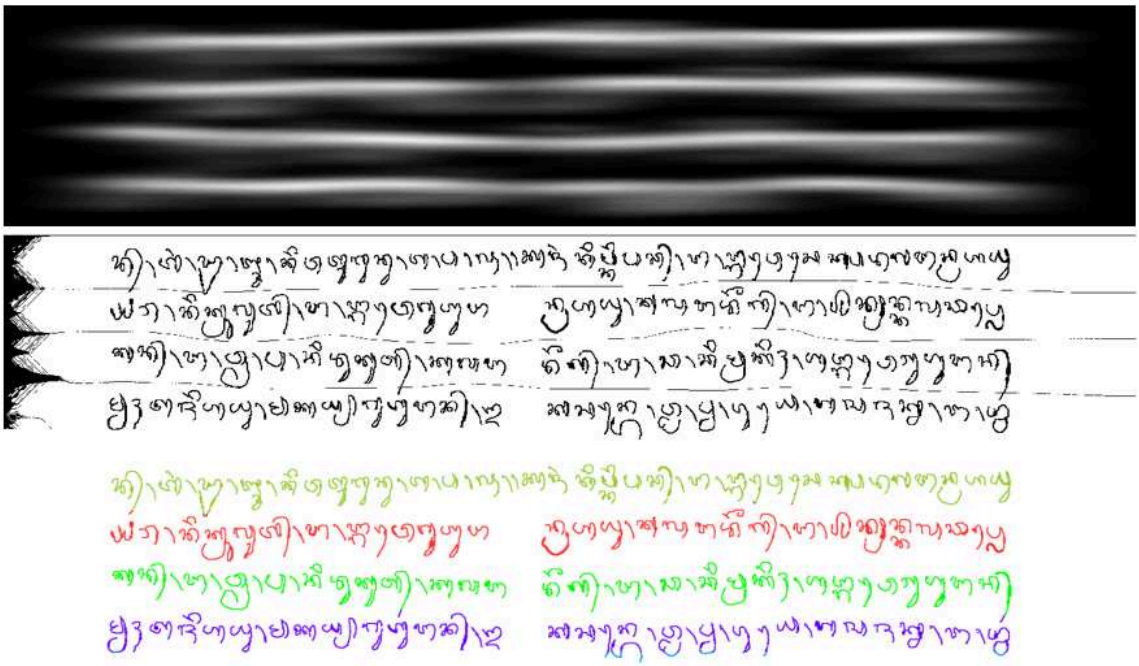


Figure 4.13: From top to bottom: Example of results: image of the energy function, the text line segmentation path generated from shredding method, and the text line segmentation result

one favouring the path that stays close to the estimated medial lines (vertical distance cost function V). Figure 4.14 illustrates an example of an optimal path.

4.3.2.3 Conclusions

The segmentation of the document image into text line segments is needed prior the text recognition step. In this section, some text line segmentation methods that work on binary images were presented. These methods take advantage of the possibility to analyze the connected components of the binary images. For example to estimate the average size of the letters or characters that can help in defining the size of the blurring window for the shredding method and the column size for the projection profile method.

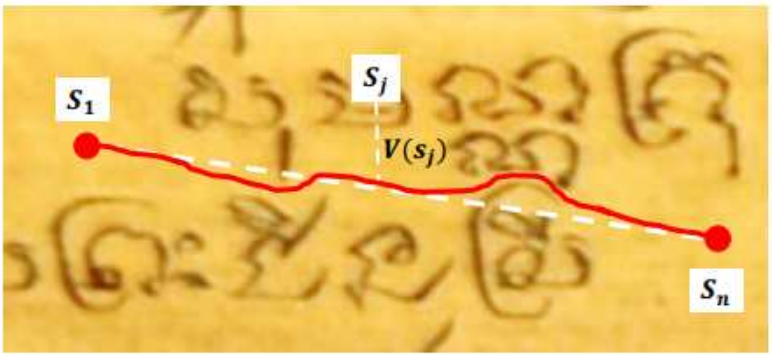


Figure 4.14: An example of an optimal path going from the start state S_1 to the goal state S_n

The drawback of this approach comes from the fact that in some cases the binarization is not optimal. The projection profile based method does not work well for curved or slanted text lines. Some methods that can be directly applied to grayscale images were also described. These methods apply some transformations or energy functions from the image intensity to get the preliminary information about the text lines for further segmentation processing steps.

4.3.3 Isolated Glyph Recognition

In a DIA system, the word or text recognition task are generally categorized in two different approaches: the segmentation based and the segmentation free methods. In the case of segmentation based method, the isolated character recognition task is a very important process [36]. Isolated handwritten character recognition (IHCR) has been the subject of intensive research during the last three decades. Some IHCR methods have reached a satisfactory performance especially for Latin script. However, development of IHCR methods for other various new scripts remains a major task for researchers. For example, the IHCR task for historical documents discovered in the palm leaf manuscripts.

4.3.3.1 Handcrafted Feature Extraction Methods

Usually, an IHCR system consists of two main steps: feature extraction and classification. The performance of an IHCR system greatly depends on the feature extraction step. The goal of feature extraction is to extract the most suitable information from raw data for classification purpose [36]. Choosing efficient and robust feature extraction methods plays a very important role to achieve high recognition performance in an IHCR and OCR [36]. A proper feature extraction and a correct classifier selection can increase the recognition rate [86]. The performance of the system depends on a proper feature extraction and a correct classifier selection [86].

Since the beginning period of pattern recognition research, many feature extractions methods for character recognition have been presented in the literature. Many feature extraction methods have been proposed to perform the character recognition task [36, 53, 54, 87, 88, 86, 89, 90, 91]. Each feature extraction method has its own advantages or disadvantages over other methods. In addition, each method may be specifically designed for some specific problem. Most of feature extraction methods, extract the information from binary image or grayscale image [87]. Some surveys and reviews on features extraction methods for character recognition were already reported [92, 93, 94, 95, 96, 97].

These methods have been successfully implemented and evaluated for recognition of Latin, Chinese and Japanese characters as well as digit recognition. However, only few systems are available in the literature for other Asian scripts recognition. For example, some of the works are for Devanagari script [33, 87], Gurmukhi script [36, 37, 35, 34], Bangla script [86], and Malayalam script [39]. Those documents with different scripts and languages surely provide a new research problem, not only because of the different shapes of characters, but also because of variance in the writing style of each script: shape of the characters, character positions, separation or connection between the characters in a text line.

In this sub section, we describe some most commonly used features for character recognition: projection histogram [39, 87, 86], celled projection [86], distance profile [39, 87],

crossing [87, 86], zoning [39, 54, 87, 88], moments [33, 39], some directional gradient based features [36, 89], Kirsch Directional Edges [87], and Neighborhood Pixels Weights (NPW) [87].

4.3.3.1.1 Projection Histogram

The projection histogram was used for the first time in to recognize the Arabic numerals with photoelectric scanners to produce electrical output in a hardware based OCR [98]. In this hardware, the black portions of characters modulate the output of a column of photocells which compose the scanner. Pulse code combinations will be generated based on the total number of black pulses per scan and the number of long-black pulses per scan.

Projection histogram counts the total number of black pixels (as character part or foreground pixels) in each row (horizontal histogram) or in each column (vertical histogram) of an image. The projection histogram can also be calculated in diagonal directions (left or right diagonal wise) [39]. The horizontal projection histogram can be represented as

$$H_i = \sum_j f(i,j) \tag{4.14}$$

while the vertical projection histogram can be represented as

$$V_j = \sum_i f(i,j) \tag{4.15}$$

where *f(i,j)* is the pixel value of ith row and jth column of the image, and the foreground pixel is considered as 1.

The vertical projection histogram will be slant invariant whereas the horizontal projection histogram is not. The width of character strokes can be estimated by using this histogram projection, but this feature can not provide any information about stroke width variations. This method can also be used as a segmentation method to segment text lines, words and characters [87].

4.3.3.1.2 Celled Projection

Celled Projection (CP) is a rapid feature extraction method proposed by Hossain et al [86]. This feature computes the projection of each section or partition of an image. For an image of size *m x n*, first, the image is partitioned into *k* horizontal regions, and then the projection histogram is calculated for each region. For a horizontal celled projection, the feature vector of the rth cell (or region) can be represented as

$$\begin{aligned} P_r &= \langle p_1, p_2, ..., p_m \rangle \\ p_i &= \sum_{j=1}^{n/k} f(i, \frac{n(r-1)}{k} + j) \end{aligned} \tag{4.16}$$

where *f(x,y)* is the pixel value of the xth row and yth column of the image, and the foreground pixel is considered as 1. The complete feature vector for the image is given by

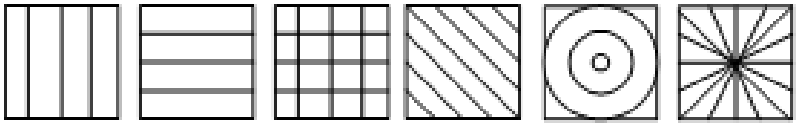


Figure 4.15: Type of Zoning (from left to right: vertical, horizontal, block, diagonal, circular, and radial zoning) [14]

$$V = P_1 \cup P_2 \cup ... \cup P_k. \tag{4.17}$$

The celled projection can also be calculated in the vertical and diagonal directions. The celled projection had been tested to recognize the Bangla handwritten numerals, and the results show a good recognition accuracy [86].

4.3.3.1.3 Distance Profile

The distance profiles extract the distance from the outermost foreground pixel to the boundary of the image [87]. The distance can be traced in horizontal, vertical and radial direction [39]. By using only horizontal and vertical directions, this feature provides four distance profiles from four sides: left, right, top and bottom. This feature gives the information about the structure of the outer contour of the character image, but without giving any information about the interior structure such as loops or number of strokes. If it is not necessary to calculate the distance profiles in each row or column, the distance profiles can be selected only in some specific rows and columns [87].

4.3.3.1.4 Crossing

Crossing counts the number of pixel transitions from background to foreground or vice versa along each straight line on the image [86]. Horizontal crossing is calculated on the line along the rows, and vertical crossing is calculated on the line along the columns [87]. This feature gives the information about the number of strokes on one line of the image, but it does not give any information about the width of strokes. Crossing can be calculated from the original character image or by first skeletonizing the character image.

4.3.3.1.5 Zoning

Zoning is computed by dividing the image into N smaller zones: vertical, horizontal, square, diagonal left and right, radial or circular zone (see Figure 4.15). The local properties of image are extracted on each zone. Zoning can be implemented for binary image and grayscale image [87]. For example, in a binary image, the percentage density of character pixels in each zone is computed as a local feature [88]. In a grayscale image, the average of gray value in each zone is considered as a local feature [39]. Zoning can be easily combined with other feature extraction methods [86] such as all types of pixel profile contour, density and area based feature. For example in [54], the features based on the direction of line segments within a character image were calculated in zoned areas of equal size.

4.3.3.1.6 Moments

Moments represent pixel distribution around the center of gravity of the character and it provides a very effective global information about the character shape [39]. Invariant

moments are invariant features under translation, scaling, rotation and reflection [33]. Moments generate discriminative features to describe the particular pattern characteristics. Any ordinary geometrical moment is calculated with

$$m_{pq} = \sum_X \sum_Y X^p Y^q f(X,Y) \tag{4.18}$$

where p,q=0,1,2,... are the order of the moments to be calculated.
The components of the centroid of moments are calculated with

$$\bar{X} = \frac{m_{10}}{m_{00}} \tag{4.19}$$

and

$$\bar{Y} = \frac{m_{01}}{m_{00}}. \tag{4.20}$$

The central moment is calculated with

$$\mu_{pq} = \sum_X \sum_Y (X-\bar{X})^p (Y-\bar{Y})^q f(X,Y). \tag{4.21}$$

Scale invariant moments using central moments of up to order 3 are computed with

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^\gamma} \tag{4.22}$$

where

$$\gamma = \frac{(p+q)}{2} + 1 \tag{4.23}$$

for p+q=2,3,...
The 7 invariant Hu's moments using scale invariant moment matrix of order 3 are computed with

$$\phi_1 = \eta_{20} + \eta_{02} \tag{4.24}$$

$$\phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \tag{4.25}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{4.26}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{4.27}$$

$$\begin{aligned}\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{4.28}$$

$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \tag{4.29}$$

$$\begin{aligned}\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2].\end{aligned} \tag{4.30}$$

Moment based features should be performed correctly using floating point arithmetic operations. The invariant moments can be evaluated by using log of the absolute value of the moment.

4.3.3.1.7 Directional Gradient Based Features

In a 2D image, the gradient vector at a point is the direction of largest possible intensity increase at that point [36]. The rate of change in that direction is the magnitude of the gradient at that point itself. The gradient is computed by calculating the derivatives in the vertical and horizontal directions. There are many operators that can be used to compute the gradient, such as the Sobel operator, Robertz operator or Prewitt operator. For example, the horizontal gradient component $\mathbf{G_x}$ and the vertical gradient component $\mathbf{G_y}$ of an image I by using Sobel operator and 8-neighborhood of pixel *(i,j)* are computed with

$$\begin{aligned} G_x(i,j) = I(i-1,j-1) + 2 \times I(i-1,j) + I(i-1,j+1) - \\ I(i+1,j-1) - 2 \times I(i+1,j) - I(i+1,j+1) \end{aligned} \tag{4.31}$$

and

$$\begin{aligned} G_y(i,j) = I(i-1,j-1) + 2 \times I(i,j-1) + I(i+1,j-1) - \\ I(i-1,j+1) - 2 \times I(i,j+1) - I(i+1,j+1). \end{aligned} \tag{4.32}$$

By applying the Robertz filter and a 2x2-neighborhood of the pixel, the gradient components are computed with [89]

$$G_x(i,j) = I(i+1,j+1) - I(i,j) \tag{4.33}$$

and

$$G_y(i,j) = I(i+1,j) - I(i,j+1). \tag{4.34}$$

The magnitude of the gradient is then computed with

$$|G(i,j)| = \sqrt{(G_x(i,j))^2 + (G_y(i,j))^2}. \tag{4.35}$$

The direction of the gradient can be computed with

$$\Theta(i,j) = tan^{-1} \frac{G_y(i,j)}{G_x(i,j)}. \tag{4.36}$$

The directional gradient based features are formed by accumulating the magnitude of the gradient along a different direction range as explained in the following steps [36, 89].

1. Quantization of the direction of gradient into 32 levels with $\pi/16$ interval.
2. The image is normalized and then it is divided into 81 blocks (9 horizontal and 9 vertical). In each block, the magnitude of the gradient is accumulated based on the 32 quantized direction of gradient. This step produces 81 local spectra of direction.
3. Down sampling of spatial resolution from 9 x 9 to 5 x 5 with 5 x 5 Gaussian filter.

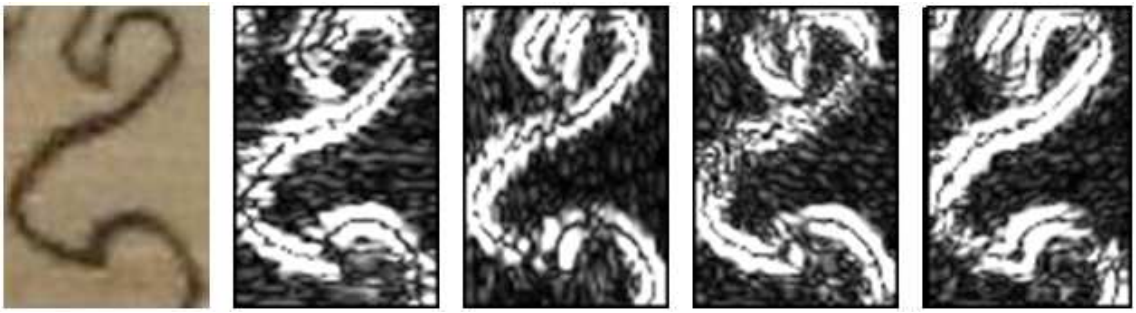


Figure 4.16: Four directional Kirsch edge images

4. Down sampling of directional resolution from 31 to 16 with a weight vector [1 4 6 4 1]^T

5. After this step, a feature vector of size 400 is produced (5 x 5 spatial x 16 directional resolution). In [36], the second down sampling of directional resolution from 16 to 8 with a weight vector [1 2 1]^T is applied, so it produces the final feature vector of size 200 (5 x 5 spatial x 8 directional resolution).

6. A transformation of y=$x^{0.4}$ is applied to make the distribution of the features Gaussian-like.

The Gaussian filter and the weight vectors are used as the hight cut filters to reduce the aliasing effect after the down sampling step.

4.3.3.1.8 Kirsch Directional Edges

The Kirsch edges method is a non-linear edge enhancement [87]. Let A_i (i=0,1,2,...,7) be the eight neighbors of the pixel (x,y), i is taken as modulo 8, starting from the top left pixel at the moving clock-wise direction. Four directional edge images are generated (Figure 4.16) by computing the edge strength at pixel (x,y) in four (horizontal, vertical, left diagonal, right diagonal) directions, defined as G_H, G_V, G_L, G_R, respectively [87]. They can be denoted as

$$G_H(x,y) = max(|5S_0 - 3T_0|, |5S_4 - 3T_4|) \tag{4.37}$$

$$G_V(x,y) = max(|5S_2 - 3T_2|, |5S_6 - 3T_6|) \tag{4.38}$$

$$G_R(x,y) = max(|5S_1 - 3T_1|, |5S_5 - 3T_5|) \tag{4.39}$$

$$G_L(x,y) = max(|5S_3 - 3T_3|, |5S_7 - 3T_7|) \tag{4.40}$$

where S_i and T_i can be computed by:

$$S_i = A_i + A_{i+1} + A_{i+2} \tag{4.41}$$

$$T_i = A_{i+3} + A_{i+4} + A_{i+5} + A_{i+6} + A_{i+7}. \tag{4.42}$$

Each directional edge image is thresholded to produce a binary edge image. The binary edge image is then partitioned into N smaller regions. Then, the edge pixel frequency in each region is computed to produce the feature vector.

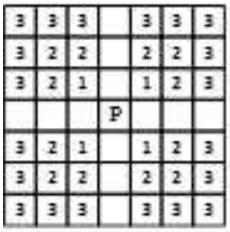


Figure 4.17: Neighborhood pixels for NPW features [14]

4.3.3.1.9 Neighborhood Pixel Weights (NPW)

Neighborhood Pixels Weight (NPW) was proposed by Satish Kumar [87]. This feature may work on binary as well as on gray images. NPW considers four corners of neighborhood for each pixel: top left, top right, bottom left, and bottom right corner. The number of neighbors considered on each corner is defined by value of layer level (see Figure 4.17). Level 1 considers only pixels in layer 1 on each corner (1 pixel), level 2 considers pixels in layer 1 and 2 (4 pixels), and level 3 considers pixels in all layers (9 pixels). In the case of a binary image, the weight value on each corner is obtained by counting the number of character pixel, divided by total number of neighborhood pixels on that corner. For grayscale image, the weight value on each corner is obtained by summing up the gray level of all neighborhood pixels, divided by the maximum possible weight due to all neighborhood pixels on that corner (nb_neighborhood_pixels x 255). Four weighted plans are constructed for each corner from the weighted value of all pixels on the image. Each plane is divided into N smaller regions, and the average weight of each region is computed. The feature vector is finally constructed from the average weight of each region from each plane (N x 4 vector dimension).

4.3.3.1.10 Histogram of Gradient (HoG)

The gradient is a vector quantity comprised of magnitude as well as directional component computed by applying its derivatives in both horizontal and vertical directions [36]. The gradient of an image can be computed either by using for example a Sobel, Roberts or Prewitt operator. The gradient strength and direction can be computed from the gradient vector. The gradient feature vector used in [36] is formed by accumulating the gradient strength separately along different directions.

To compute the histogram of gradient (HoG), first, we calculate the gradient magnitude and gradient direction of each pixel of the input image. The gradient image is then divided into some smaller cells, and in each cell, we generate the histogram of directed gradient by assigning the gradient direction of each pixel into a specific range of orientation bin which are evenly spread over 0 to 180 degrees or 0 to 360 degrees (Figure 4.18 & 4.19). The histogram cells are then normalized with a larger overlap connected blocks.

The final HoG descriptor is then generated from all concatenated vectors of the histogram after the block normalization process.

4.3.3.2 Convolutional Neural Network (CNN)

The CNN is different with the standard feedforward neural networks in terms of number of connections and parameters for layers with similar size. CNN use much fewer

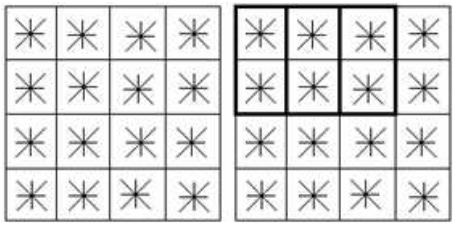


Figure 4.18: An image with 4x4 oriented histogram cells and 2x2 descriptor blocks overlapped on 2x1 cells

Figure 4.19: The representation of the array of cells HoG [14]

connections and parameters [99], but the capacity of a CNN can still be designed by the depth and breadth of the layers [100]. CNNs started to be widely used in 1990 but it is really becoming the leading method for supervised learning after the winning event for the ImageNet challenge [100]. Nowadays, CNN is applied to many tasks especially in classification problems. CNN consists of multilayer with each layer performing some specialized functionality as follows:

- Convolutional layers. The convolutional layers play the main role for training. A feature extraction step is not required in a CNN. The network considers the value of each pixel of the input image as the input layer or the gray intensity can be considered as the feature.
- Rectifier layers for non linearity. The function $f(x)=max(0,x)$ or Rectified Linear Unit (ReLU) layers are often used. Because it is a non-saturating nonlinearity and faster, while function tanh $f(x)=tanhx$ or sigmoid $f(x) = (1+e^{-x})^{-1}$ are traditional saturating neuron models [100].
- Pooling layers for reducing the number of inputs. The max-pooling layers are often used.
- The classification is performed by the fully connected layers. After training the CNN, instead of performing classification using the fully connected layers, one can feed features from the last convolution layer into an SVM classifier [99]. This last layer calculates the loss function of the network.

The convolutional-rectifier-pooling layers can be stacked to build a deeper CNN. The multilayer convolutional neural networks (CNN) have been proven to be very effective in areas such as image recognition and classification. CNNs produce a collection of feature map which are learned in small but deep spatial area. They process images bottom-up and are trained discriminatively and purely supervised. CNNs need more computational

power and time than training a SVM [99] and to reduce overfitting in the fully-connected layers, a regularization method called "dropout" is very effective to be employed [100].

4.3.3.3 Unsupervised Feature Learning (UFL)

Unavailable transcribed data training is one of the main problems in historical text recognition [29]. It is widely known that CNNs should be trained with enough labeled training data to avoid severe overfitting [100]. But the labeling or annotating process for the training data is a very time consuming task and very costly [29, 51]. By doing it manually by a human, the human error and the inconsistency of label decision often happens. In the beginning, a CNN need a large amount of labeled data to model the supervised learning. Now, it can be combined with both supervised and unsupervised methods [99]. Unsupervised feature learning is a method to learn features from unlabeled data. UFL tries to build a good feature representations from unlabeled input data for higher level tasks such as classification [101, 102].

To learn features from images, the common framework for UFL is as follow [101].

1. Extract random patches from the unlabeled training images. If the size of each patch image is w x w, then w is the size of the receptive field. Each patch image can be represented as a vector in $\mathbb{R}^N$ of pixel intensity values, with $N = w$ x w x d, where d is the number of channels of patch images. A set of vectors from m patch images will be represented as $X = \{x^{(1)}, ..., x^{(m)}\}$, where $x^{(i)} \in \mathbb{R}^N$.

2. Do the pre-processing steps for the collection of patch images. Pre-processing steps will be applied to X. One common pre-processing step is the normalization of every patch $x^{(i)}$ by substracting the mean and dividing by the standard deviation of its elements which visually correspond to local brightness and contrast normalization. Whitening is also another optional pre-processing step after normalization [101, 102]. Whitening is a way to de-correlate the data.

3. Apply an unsupervised learning algorithm to learn features from patch images. An unsupervised learning algorithm can now be applied to X. An unsupervised learning algorithm can be formally defined as an algorithm that takes the dataset X as input, and outputs a function $f : \mathbb{R}^N \rightarrow \mathbb{R}^K$. This function maps each input vector $x^{(i)}$ in X to a new feature vector of K features, where K is a parameter of the unsupervised learning algorithm. There are two requirements that need to be satisfied by the mapping function $f(x)$ [103]. First, there must be at least one feature that is similar for images of the same category (invariance). Second, there must be at least one feature that is sufficiently different for images of different categories (ability to discriminate). For example the K-Means clustering algorithm. K-Means clustering is a much simpler and faster method and it has been shown to achieve comparable results with other methods [102]. K-Means clustering algorithm is widely used as an alternative unsupervised learning module for building codebooks of "visual words" to define higher level image features [104, 105]. K-Means clustering learns K centroids $c^{(k)}$ from the input data X. Clusters are groups of data having a small distance to the cluster centroids. There are two variants of mapping function f for K-Means clustering. The standard K-Means clustering, called K-means hard, uses a hard-assignment coding scheme, 1-of-K, as

$$f_k(x) = \begin{cases} 1, k = argmin_j \left\| c^{(j)} - x \right\|_2^2 \\ 0, otherwise. \end{cases} \tag{4.43}$$

The second variant, called K-means triangle, is a non-linear mapping that maps the feature vectors in a softer way with

$$f_k(x) = max\{0, \mu(z) - z_k\} \tag{4.44}$$

where

$$z_k = \left\| x - c^{(k)} \right\|_2 \tag{4.45}$$

and $\mu(z)$ is the mean of the elements of z.

4. After having the unsupervised learned features, we can train a linear classifier with a set of labeled training images to predict the labels of a feature vector.

4.3.3.4 Methods from ICFHR 2016 Competition Challenge 3: Isolated Character Recognition of Balinese Script in Palm Leaf Manuscript Images

Although many standard features extraction methods for the isolated character recognition have been implemented and evaluated by using standard recognition schemes, there are still many possibilities for variations in the isolated character recognition schemes proposed by many researchers. These schemes differ not only in the selection of system parameters, but also in the pre-processing step, design, type and combination of features and classifiers. Several isolated character recognition methods proposed by researchers in this competition provide an overview of how diverse the development of the isolated character recognition system is.

Method from Group 1 (ICFHR G1): (VMQDF and VCMF method) The input image is first processed through the following steps: 1) the input image is resized to 64x64 by linear normalization and a median blur with a window width of 3 is applied on the normalized image. The resulting image is denoted as I1, 2) the binary image Ib of I1 is computed by the OTSU method, 3) I1 is then processed by the pre-processing method described in [66] to overcome the gray scale variation among different images. The mean and scatter employed in the experiment is set as 100 and 100, respectively. The resulting image is denoted as In. 97 virtual samples are generated based on In by the method proposed in [106]. Gradient feature is extracted using NCGF [66] for each original or virtual sample. MQDF classifier is trained based on the expanded training set. At the test stage, for each input image, 97 virtual images are generated and features are extracted in the same way. VMQDF classifies the input image by the voting of the MQDF results of the input image and all the virtual images. VCMF classifies the input image by the voting of the MQDF+CMF results of the input image and all the virtual images.

Method from Group 3 (ICFHR G3): The submitted method first segments the text pixels out of the background using the adaptive contrast method [68]. To generate more training data, they developed a synthetic data generation method based on the binary images. First, the bounding box of the character can be easily determined in the binary image.

They then crop, rotate and warp the origin images without removing the textual information of the original images. With the synthetic dataset, they build three CNN models to recognize the 133 labels. The CNN model is constructed using the convolutional network toolkit developed by Microsoft[14]. The input images are resized to 28 by 28. The first CNN network takes the origin images as network input, the second one takes the binary images as network input, and the last one combines the origin and binary images together as network input. Each CNN network has 2 convolution layers followed by two max pooling layers. After training, they further design a random forest to ensemble all the outputs of the CNN networks. The random forest is constructed using Scikit Learn package[15] with 200 trees. The final recognition result is produced by the random forest classifier.

Method from Group 5 (ICFHR G5): The finite impulse response multilayer perceptron (FIRMLP), a class of temporal processing neural networks, is a multilayer perceptron where the weights have been replaced by finite impulse response filters. The FIRMLP consists of only one type of processing unit - the finite impulse response neuron. They have extended this architecture which now consists of layers with different computational properties and recurrent connections as well. The first layer is composed of temporal quadratic units. Hence, those neurons compute polynomial kernels of order two, transforming the input space into a higher dimensional feature space. The second layer uses temporal radial basis function units and separates the extended input space by spheres. The third and fourth layer incorporate sigmoid and linear neurons, respectively, with usual FIR synapses. By using the finite impulse response filter, the neuron possess internal short time memory and represents a model for spatio-temporal processing. By adding recurrent connections, the neuron develops long time memory as well. The neural networks used as input 64x64 gray pixel images obtained as follow: 1) first, the MSRCR-Algorithms (MultiScale Retinex with Color Restoration) is applied 2) image binarization is performed by using the K-Means algorithm (from OpenImaj package) with two centers 3) and finally scaling to 64x64 pixel images is applied.

4.3.3.5 Conclusions

Isolated character recognition is an important task in the segmentation based text recognition process. The performance of segmentation based text recognition depends on proper feature extraction and classification. In this sub section, some handcrafted feature extraction methods were described. The methods that work on binary images aim to extract the inner and outer profile of the character. The methods that can be applied on grayscale images are taking into account the information about image intensity in some smaller area. They define the weighted value of pixel combinations or they calculate the direction of the intensity gradient in each cell area. The CNN proposes a learning based recognition method where the feature extraction step is not required. A CNN directly learns the feature map in small but deep spatial area. A CNN uses more computational power than a SVM but a CNN tends to produce the overfitting problem. The UFL tries to overcome the drawback of a CNN by proposing a framework to learn features from unlabeled data.

[14]CNTK http://www.cntk.ai/
[15]Scikit-Learn http://scikit-learn.org/

4.3.4 Text Transliteration with Long Short Term Memory Network (LSTM)

In many DIA systems, word or text recognition is the final task on the processing pipeline. In order to make the palm leaf manuscripts more accessible, readable and understandable to a wider audience, an optical character recognition (OCR) system should be developed. But, normally in Southeast Asian script, the speech sound of the syllable change related to some certain phonological rules. In this case, an OCR system is not adequate. Therefore, a transliteration system should also be developed to help to transliterate the ancient scripts on these manuscripts. The transliteration system will be very helpful in reading the Balinese palm leaf manuscripts for most of the young scholars who are not familiar with the Balinese script.

By definition, transliteration is defined as the process of obtaining the phonetic translation of names across languages [107]. Transliteration involves rendering a language from one writing system to another[16]. In [107], the problem is stated formally as a sequence labeling problem from one language alphabet to other. It will help to index and to access the content of the manuscripts quickly and efficiently. Many transliteration models have been proposed [107, 108, 109, 110].

In this dissertation, two approaches for text transliteration are proposed. The first approach is segmentation based transliteration. For this approach, an implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscript are presented. In this system, a rule-based engine [110] for performing transliterations is proposed. This model is based on phonetics which are based on traditional linguistic study of Balinese transliteration. A complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts is also proposed. The segmentation based approach for transliteration of Balinese script will be presented and be described in Chapter 6.

In this sub section, we will describe the second approach, a segmentation free LSTM based method for text recognition that is used and is evaluated to transliterate the words and the text lines from palm leaf manuscripts. It is popularly known that Long Short-Term Memory (LSTM) type learning network is widely used in sequence analysis problems. Text recognition and transliteration processes are typically among those problems.

In the last decade, the sequence analysis based method that uses Recurrent Neural Network - Long Short-Term Memory (RNN-LSTM) type learning network became a very popular method among the researchers in text recognition. The RNN-LSTM based method combined with a Connectionist Temporal Classification (CTC) works as a segmentation free learning based method to recognize the sequence of characters in a word or text without any handcrafted feature extraction method. RNN-LSTM is a fully supervised learning method. The target ground truth of each image train should be provided. The CTC layer will allign the output from the network with the target ground truth data by maximizing the expectation over all possible alignments. To calculate this, the forward-backward algorithm can be used efficiently.

A Recurrent Neural Network (RNN) is a modified feedforward neural network with self-connected neurons at the hidden layers. With these self-connected neurons, a RNN is

[16] https://www.accreditedlanguage.com/2016/09/09/what-is-transliteration/

able to remember the previous cell state to learn sequences in different time steps. RNN is basically an extended version of the basic feedforward neural network. In an RNN, the neurons in the hidden layer are connected to each other. RNNs offer a very good context-aware processing to recognize patterns in a sequence or time-series. One draw back of the RNN is the exploding/vanishing gradient problem. During the training process, the gradients become very large or very small. To deal with this problem, the LSTM architecture was then introduced.

LSTM has two main properties, context sensitive learning and good generalization [111, 29]. LSTM has been widely applied to text recognition task without using handcrafted features and language modeling. The raw image pixels can be sent directly as the input to the learning network and there is no requirement to segment the training data sequence. The LSTM architecture is widely known as a generic and language independent text recognizer [112]. LSTM has been used to recognize printed text and handwritten text successfully [29]. LSTM is a modified architecture of RNN. The LSTM network adds the multiplicative gates and additive feedback. A LSTM cell can forget the information inside the cell that is no longer required, and can add new information to be stored inside the cell. A LSTM cell consists of:

- Forget Gate with the Logistic Sigmoid Function. It decides which inputs to retain at any given time step. It produces a value between 0 (if the cell forgets all information inside the cell) or 1 (if the cell retains all information inside the cell).
- Input Gate. It determines which input is processed by the cell.
- Output Gate. It determines what output the cell has.

The forward training process of a simplified LSTM are as follows [51, 29]:

$$f_t = \sigma(W_{xf}.x_t + W_{hf}.h_{t-1} + b_f) \tag{4.46}$$

$$i_t = \sigma(W_{xi}.x_t + W_{hi}.h_{t-1} + b_i) \tag{4.47}$$

$$v_t = tanh(W_{xv}.x_t + W_{hv}.h_{t-1} + b_v) \tag{4.48}$$

$$o_t = \sigma(W_{xo}.x_t + W_{ho}.h_{t-1} + b_o) \tag{4.49}$$

$$C_t = f_t * C_{t-1} + i_t * v_t \tag{4.50}$$

$$h_t = o_t * tanh(C_t) \tag{4.51}$$

where

- f_t is the forget gate layer, i_t is the input gate layer, and o_t is the output gate layer.
- C_t is the Cell State, h_t is the cell output and h_{t-1} is the cell output at the previous timestep.
- b_y, $y \in \{f,i,v,o\}$ is the bias unit for the forget gate, input gate, input squashing and output gate respectively.

- W_{ij} is the weight connection between the external i^{th} and j^{th} node, e.g., W_{xf} is the weight connection between the external input and the forget gate.
- σ is the logistic sigmoid function given by

$$\sigma = \frac{1}{1 + exp(-x)}. \tag{4.52}$$

- *tanh* is the tangent hyperbolic function given by

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}. \tag{4.53}$$

The backward process of LSTM is a back propagating process of error between the output produced in forward process compared to target output. The learning weights and biases of the network will be updated during this backpropagated error learning process by using a modified backpropagation algorithm called backpropagation through time (BGT) [29].

Segmentation based and segmentation free methods can be combined to increase the performance of OCR system [51]. Some schemes that combine the strength of segmentation based and segmentation free approaches were also already reported, for example the OCRoRACT [51] and the anyOCR [29].

OCRoRACT uses a segmentation based OCR method to train on individual symbols and then uses the semi-corrected recognized text lines as the ground truth data for the segmentation-free method [51].

The anyOCR uses an unsupervised clustering algorithm as their segmentation-based OCR approach. The unsupervised clustering algorithm does not need transcribed training data. The anyOCR framework is designed to be adapted for documents with no or very limited ground truth training data. AnyOCR was reported with a better performance compared to the OCRoRACT system which utilizes Tesseract[17]. But, it is also stated that anyOCR could not overcome the general weakness when dealing with scripts like Arabic and Devanagari, where character/ligature segmentation is not trivial [29].

In this work, the OCRopus[18] framework is used to test and evaluate the transliteration task for the palm leaf manuscript collection. OCRopus is a segmentation-free open source document analysis and recognition system based on LSTM neural networks and CTC layer. OCRopy provides the functional library of the OCR system by using a contemporary LSTM architecture[19] [111]. OCRopus was already tested and evaluated on some printed historical documents such as historical Greek Polytonic scripts [113], printed English and Fraktur [111], and medieval Latin script [29, 51]. In OCRopus, the height of the image is first normalized for all inputs based on Gaussian filtering and affine transformation. The implementation of LSTM in OCRopus uses a sliding window of size 1-pixel wide. Each pixel value will be sent to one LSTM-unit in the input layer. We evaluated the dataset with unidirectional LSTM and the Bidirectional Long Short-Term Memory (BLSTM) architecture. The Bidirectional LSTM is a LSTM achitecture with two directional (forward and backward) context processing.

[17]https://github.com/tesseract-ocr/
[18]https://github.com/tmbdev/ocropy
[19]http://graal.hypotheses.org/786

4.4 Conclusions

Ground truth construction is a very important part in the development of DIA system. The ground truthed dataset is not only needed to provide a training dataset for some supervised learning DIA methods, but also to provide the evaluation and benchmark dataset to quantitatively measure the performance of the DIA methods. The nonexistence of a ground truth dataset can be a great obstacle in DIA system development. But the ground truthing processes are often very costly in term efforts, times and energies. The semi-automatic process is proposed to economize the manual efforts in dataset ground truthing step. In the real condition of implementation, even though the semi-automatic tools and schemes are already used to help the ground truthing process, manual validation by some experts in the special domain is still need.

Many binarization methods have been reported. These methods have been tested and evaluated on different types of document collections. The robustness of these methods still have to be tested and evaluated on document image collections with new specific characteristics like palm leaf manuscript images. As common starting point for the document image analysis pipeline, an optimal binarization method for palm leaf manuscripts may help in supporting the initial step of ground truth dataset construction process and in the following steps of glyph recognition and text recognition.

Text line segmentation is a crucial pre-processing step of most DIA pipelines. The task aims at extracting and separating a text region into individual lines. In this case, extracting the text lines in a document properly will ease the extraction of smaller size entities of the document, such as the words or the characters. Consequently, the performance of the OCR system is greatly influenced by the outcome of the segmentation process. In this section, some promising text line segmentation methods for palm leaf manuscript images have been presented. The methods that work on binarized images as well as the methods that can be applied on grayscale images are investigated. Each method has a different behavior to be properly applied to palm leaf manuscript images.

Although many methods for isolated character recognition have been widely developed and tested especially for Latin based scripts and alphabets, there is still a need for in-depth evaluation of those methods to be applied for various other types of scripts with optimal performance. It includes the isolated character recognition task for many South-east Asian scripts, and more specifically the scripts which were written on the ancient palm leaf manuscripts. A IHCR system is one of the most demanding systems which has to be developed for the collection of palm leaf manuscripts. Using an IHCR system will help to transliterate these ancient documents and translate them to the current language.

Considering the importance of the contents of the palm leaf manuscripts which represent an important cultural heritage, some efforts had been done to digitize these manuscripts and then to make the palm leaf manuscripts more accessible, readable and understandable for a larger public. To achieve that goal, an optical character recognition (OCR) system and a transliteration system should be developed to convert the original script into the Latin script. Those systems will help to transliterate these ancient documents, and to index and to access quickly and efficiently the content of the manuscripts. For example in the case of the Balinese palm leaf manuscripts from Indonesia, the transliteration system is the most demanding system which has to be developed and it will be very helpful

in reading these manuscripts for most of the young scholars who are not familiar with the Balinese script. It is popularly known that Long Short-Term Memory (LSTM) type learning network is widely used in sequence analysis problems. Text recognition and transliteration process are typically among those problems. A LSTM based method combined with a Connectionist Temporal Classification (CTC) works as a segmentation free learning based method to recognize the sequence of characters in a word or text without any handcrafted feature extraction method. The challenge for the transliteration of the Balinese script comes from the fact that it is a syllabic script and that the speech sound of the syllable change related to some certain phonological rules. In addition, with a very limited training data availability, some adaptations of LSTM in transliteration training scheme need to be designed, to be analyzed and to be evaluated to ensure the robustness of the results. In the next chapter, the corpus and ground truth dataset of Balinese palm leaf manuscripts which are used for all research works in this dissertation will be presented. The protocols design and the complete process from the manuscript digitization process until the dataset annotation process will be described.

Chapter 5

Corpus and Ground Truth Dataset Construction of Balinese Palm Leaf Manuscripts

This chapter presents the corpus and ground truth dataset of Balinese palm leaf manuscripts which are collected, constructed and used for all research works in this dissertation. This chapter describes the protocols design and the complete process from the manuscript digitization process until the dataset annotation process. The additional corpus and dataset from Khmer and Sundanese palm leaf manuscripts are also presented in this chapter.

5.1 Corpus of Balinese Palm Leaf Manuscripts

The corpus of palm leaf manuscripts which was collected are the sample images of the palm leaf manuscripts from Bali, Indonesia. It is hard to estimate the size of the whole collection of palm leaf manuscript in Bali, because most of the palm leaf manuscript collections are kept by the private families as a private collection. For this research, in order to obtain the variety of the manuscript images, the sample images were collected from 23 different collections (contents), which came from 5 different locations (regions): 2 museums and 3 private families (Figure 5.1). They consist of randomly selected 10 collections from Museum Gedong Kirtya, City of Singaraja, Regency of Buleleng, North Bali, Indonesia, 4 collections from manuscript collections of Museum Bali, City of Denpasar, South Bali, 7 collections from the private family collection from Village of Jagaraga, Regency of Buleleng, and 2 others collections from private family collections from the Village of Susut, Regency of Bangli and from Village of Rendang, Regency of Karangasem. From those 23 collections, we captured 393 pages of palm leaf manuscript. A summary of the collection is listed in Table 5.1.

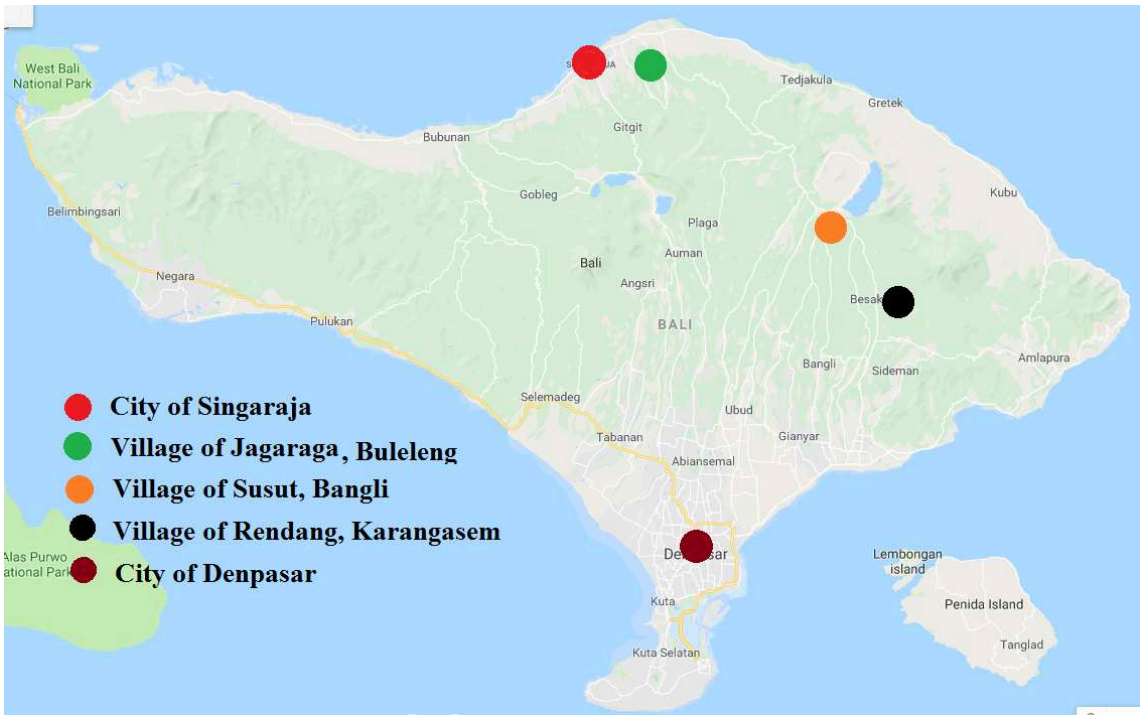


Figure 5.1: Five different locations of palm leaf manuscript corpus in Bali

5.2 Digitization Process

To capture the manuscripts, a Canon EOS 5D Mark III camera was used. The camera settings are as follow [18]: F-stop: f/22 (diaphragm), exposure time: 1/50 sec, ISO speed: ISO-6400, focal length: 70 mm, flash: On - 1/64, distance to object: 76 cm, focus: Quick mode - Auto selection 'On'. We also designed a black box camera supported made of wood to avoid the irregular lighting/luminance condition and to fit our semi outdoor capturing location (Figure 5.2 and 5.3). This camera support was intendedly designed to be used under several restricted conditions given by the museum or the owner of the

Table 5.1: Corpus collection of palm leaf manuscripts from Bali, Indonesia

Location	Collection Code	Content	Nb of captured pages
Museum Gedong Kirtya, City of Singaraja (10 collections)	IIA-10-1534	Awig-awig Desa Tunju	8
	IIA-5-789	Sima Desa Tejakula	8
	IIB-2-180	Dewa Sasana	8
	IIIB-12-306	Panugrahan Bhatara Ring Pura Pulaki	8
	IIIB-42-1526	Buwana	8
	IIIB-45-2296	Pambadah	8
	IIIC-19-1293	Krakah Sang Graha	8
	IIIC-20-1397	Taru Pramana	8
	IIIC-23-1506	Siwa Kreket	8
	IIIC-24-1641	Tikas Patanganan Weda	8
Museum Bali, City of Denpasar (4 collections)	MB-AdiParwa(Purana) -5338.2-IV.a	Adi Parwa (Purana)	40
	MB-AjiGriguh -5783-107.2	Aji Griguh	20
	MB-ArjunaWiwaha -GrantangBasaII	Arjuna Wiwaha-Grantang Basa II	30
	MB-TaruPramana	Taru Pramana	40
Village of Jagaraga, Regency of Buleleng (7 collections)	JG-01	Unknown	16
	JG-02	Unknown	10
	JG-03	Unknown	16
	JG-04	Unknown	12
	JG-05	Unknown	8
	JG-06	Unknown	5
	JG-07	Unknown	10
Village of Susut, Regency of Bangli (1 collection)	Bangli	Sabung Ayam	82
Village of Rendang, Regency of Karangasem (1 collection)	WN	Surat Jual Beli Tanah	24
Total Number of Pages			393

Table 5.2: Profile of Ground Truthers

Group	Profession	Number of people	Age	Balinese script literacy	Computer literacy
Group1	Balinese philologist	2 persons	25-40 y.o.	Advanced	Basic
Group2	Bachelor student in Balinese literature	±20 persons	18-22 y.o.	Intermediate	Intermediate
Group3	Bachelor student in Informatics	±70 persons	18-22 y.o.	Basic	Advanced

manuscripts. Two additional lights were added inside the black box support with White Neon 50 cm 20 watt. Thumbnail samples of the captured images are shown in Figure 5.4. To digitize large collections of palm leaf manuscript and to place them online, the philologists rated the quality of these images are good enough.

5.3 Ground Truth Dataset Construction

In order to develop and to evaluate the performance of the document analysis methods, the dataset and the corresponding ground truth data are required. Based on our knowledge, there is no existing public dataset and ground truth image for palm leaf manuscripts. Therefore, creating a new dataset and ground truth image for palm leaf manuscripts is a necessary step for the research community.

In line with the DIA tasks that will be investigated in this research, we created multiple components of the ground truth construction for palm leaf manuscripts as follows:

- Transliterated text ground truth of manuscript collection
- Binarized images ground truth dataset construction
- Text line segmented images ground truth construction
- Word annotated images ground truth construction
- Isolated glyph annotated images ground truth construction

Table 5.2 describes the profile of ground truthers for the ground truth dataset construction in this research. Figure 5.5 shows the overall scheme of ground truth construction for all datasets of palm leaf manuscripts. The following sub sections will describe each components of the ground truth construction.

And finally, based on the results of each component of the ground truth construction, we have built eight components of Balinese palm leaf manuscript dataset as follows:

- Transliterated manuscript image ground truth dataset
- Binarized image ground truth dataset
- Text line segmented image ground truth dataset
- Word annotated image ground truth dataset

Figure 5.2: Camera support for digitizing of palm leaf manuscripts

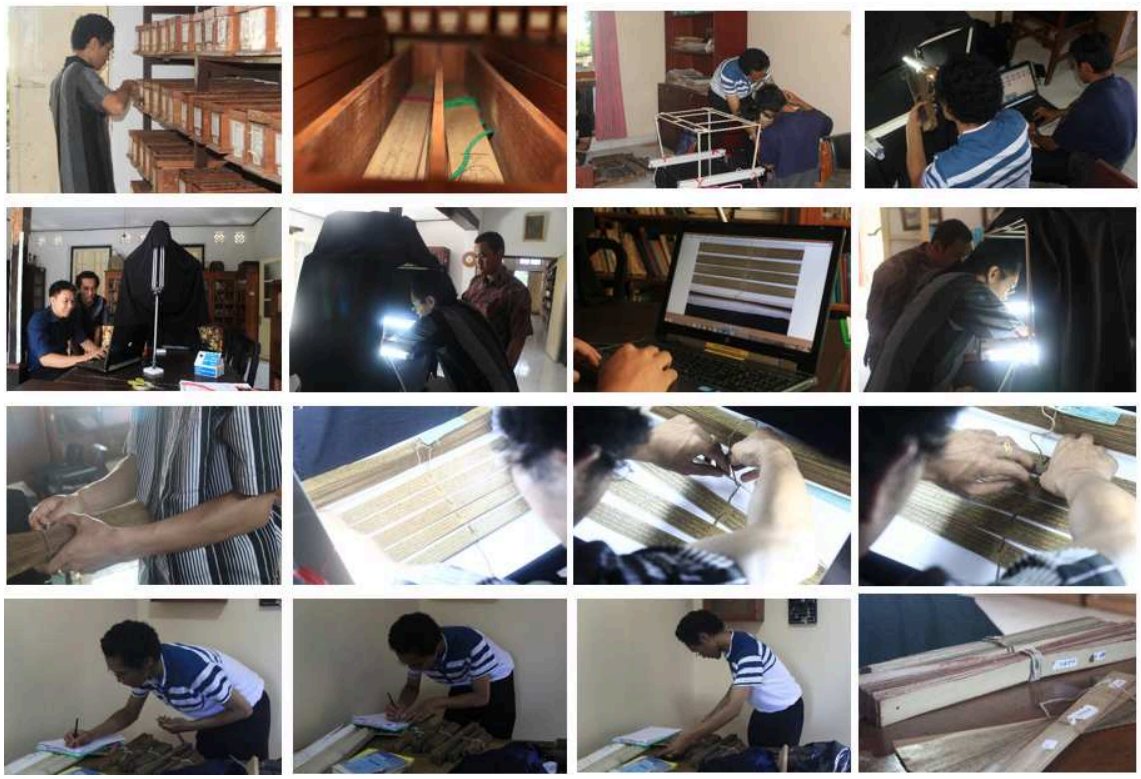


Figure 5.3: Digitization process of palm leaf manuscripts

- Query-by-example word spotting dataset
- Isolated glyph annotated image ground truth dataset
- Query-by-example glyph spotting dataset
- Page image with glyph segmentation and recognition ground truth dataset

Each of those ground truth dataset components will be described in Section 5.4.

5.3.1 Transliterated Text Ground Truth of Manuscript Collection

The ground truth of transliterated text for all manuscripts are provided and manually typed on a text editor by the Balinese philologists (Figure 5.6).

5.3.2 Binarized Images Ground Truth Dataset Construction

To evaluate the performance of the binarization method, two approaches are widely used. The first approach evaluates the binarization method based on the character recognition rate reached by an OCR system applied on those binarized images [13]. But this approach has been criticized that the binarization method is evaluated in their interaction with other process on document analysis pipeline. The second approach evaluates the binarization method by comparing the difference between binarized image and a ground truthed binarized image pixel-by-pixel [2, 114]. In the case where the OCR system for Balinese script is not available yet, the ground truth binarized image of palm leaf manuscripts has to be created to be able to quantitatively measure and compare the performance of all binarization methods.

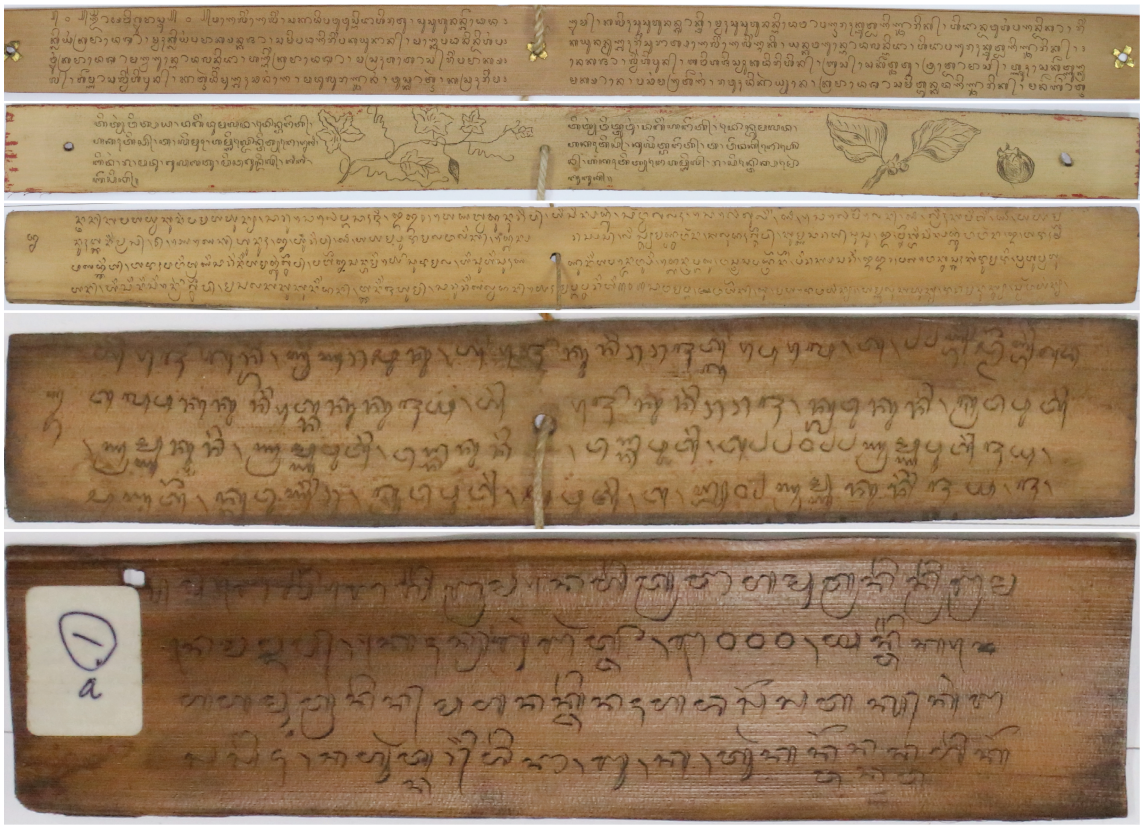


Figure 5.4: Sample images of palm leaf manuscript (top to bottom) from a) Museum Gedong Kirtya, Singaraja, b) Museum Bali, Denpasar, c) Village of Jagaraga, Buleleng, d) Village of Susut, Bangli, e) Village of Rendang, Karangasem

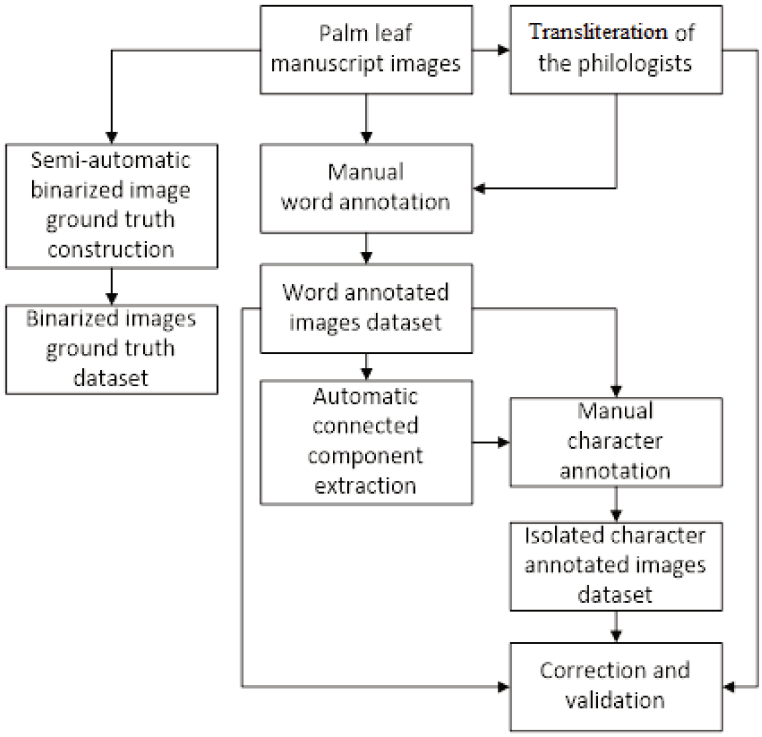
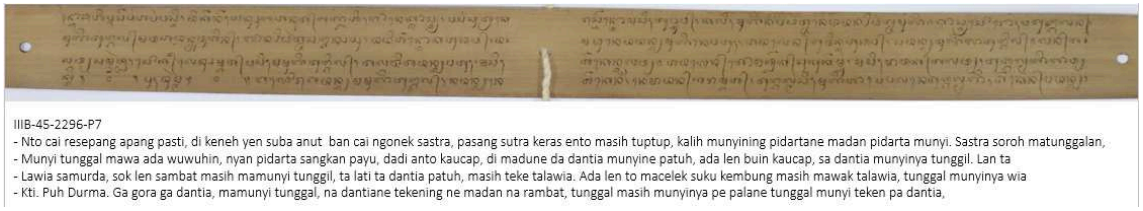


Figure 5.5: Overall scheme of ground truth dataset construction



IIIB-45-2296-P7
- Nto cai resepang apang pasti, di keneh yen suba anut ban cai ngonek sastra, pasang sutra keras ento masih tuptup, kalih munyining pidartane madan pidarta munyi. Sastra soroh matunggalan,
- Munyi tunggal mawa ada wuwuhin, nyan pidarta sangkan payu, dadi anto kaucap, di madune da dantia munyine patuh, ada len buin kaucap, sa dantia munyinya tunggil. Lan ta
- Lawia samurda, sok len sambat masih mamunyi tunggil, ta lati ta dantia patuh, masih teke talawia. Ada len to macelek suku kembung masih mawak talawia, tunggal munyinya wia
- Kti. Puh Durma. Ga gora ga dantia, mamunyi tunggal, na dantiane tekening ne madan na rambat, tunggal masih munyinya pe palane tunggal munyi teken pa dantia,

Figure 5.6: Transliterated Text Ground Truth of Manuscript

To create the binarized images ground truth dataset of palm leaf manuscripts, we have analyzed the previous works on construction of ground truth binarized images [13], especially based on the method proposed and used for DIBCO competition series [3, 114], as previously explained in Sub Section 4.2.2. The purpose of our work is to achieve better ground truth binarized images for low quality palm leaf manuscripts. In this framework, human intervention plays a very important role by performing the manual correction of the character skeletons based on character edges (detected using the Canny algorithm [63]). Based on the preliminary experiments, it is expected to obtain a good initial binarized image as the input to the next process of ground truth creation [1]. The initial binarization method used in the stage of construction of skeletonized ground truth image should be able to generate an optimal and acceptable 'good enough' skeleton which detects and keeps the form of the characters. An image of a skeleton generated in this step will facilitate the manual correction process. The more the skeleton is correct, the more the manual process is easier and faster. For a nondegraded palm leaf manuscript, a simple global thresholded binarization method is sufficient to generate an acceptable binarized image and optimal image of skeleton. However, this method is not adapted to degraded palm leaf manuscripts. Figure 5.7 shows some examples of skeletonized image generated with Matlab standard function bwmorph[1] from different binarized images using different binarization methods (see Sub Section 3.2.1 Figure 3.7). Influenced by the dried palm leaf texture, the stroke of characters in the palm leaf manuscripts is thickened and widened. As a consequence, a lot of small short unuseful branches on the skeleton are generated. Because of the poor quality of the binarized and skeletonized images, the step of manual correction of the skeleton is very time consuming. It takes almost 8 hours for only one image of palm leaf manuscript. Another important remark, superimposing the image of a skeleton on the original image to guide the manual correction process is not enough. A priori knowledge of the form of ancient characters is mandatory to guarantee that the incomplete character skeleton can be completed in a more natural trace as how the characters have been originally written. The manual correction process should be done by a philologist or at least by a person who knows well how to write the ancient characters with a guide for the transliteration of the manuscript provided by a philologist.

Therefore, in the case of degraded and low quality palm leaf manuscripts, this study focuses on the development of an initial binarization process for the construction of ground truth binarized images. The need for a specific scheme which adapts and performs better in constructing the ground truth of binarized images for palm leaf manuscripts should be analyzed to achieve a better ground truth for low quality palm leaf manuscripts. The proposed specific binarization scheme for binarized images ground truth dataset construction will be described in Sub Section 5.3.2.1.

[1] http://fr.mathworks.com/help/images/ref/bwmorph.html

Figure 5.7: Examples of image of skeleton (left to right and up to bottom) [1] generated from binarized image of Otsu [2, 3], Niblack [4, 5, 6, 7, 8], Rais [5], NICK [4]

5.3.2.1 Proposed specific binarization scheme

In order to overcome the binarization problem on degraded and low quality palm leaf manuscripts, we propose a 'semi-local' concept [1]. The idea of this method is to apply a powerful global binarization method on only a smaller and precise local character area. The binarization scheme consists of several steps as illustrated in Figure 5.8. First, the edge detection with the Prewitt operator is applied to get the initial surrounding area of the line-strokes of each character. Based on our visual observation, Prewitt leads to high edge response on the inner part of the characters, and it gives a good approximate area for the skeleton. Whereas Canny leads to a high edge response on the outer side of text stroke, and it detects over sensitively the textural part of the palm leaf background. The grayscale image of the edge is then binarized with Otsu's method to get the first binarized image of the palm leaf manuscript. A median filter is then applied to this binarized image in order to remove noise. After noise reduction, some characters might be affected and broken. A dilation process is applied to recover and reform the broken parts of the character. The method constructs the approximated character area using the Run Length Smearing (RLS) method [115]. The smearing method should be done properly, so the missing/undetected character area can be detected completely. The RLS in row wise will cover the missing area in horizontal strokes character line, meanwhile the RLS in column wise will cover the missing area in vertical strokes character line. The output of those steps is a binarized image with an approximated character area in black, and the background area in white. The next step is the main concept of this scheme. Otsu's binarization method is applied for the second time, but locally only within a limited character area, defined by each connected component from the first binarized image generated (Figure 5.9).

The first initial binarization process is needed to roughly separate text from the background, and it provides a first binary image for the skeletonizing process of the characters on manuscript. After the initial binarization process, the method finally performs a morphological-based thinning method to get the skeleton of the character. The thinned image still normally has the unwanted branchs, so it applies a morphological-based prun-

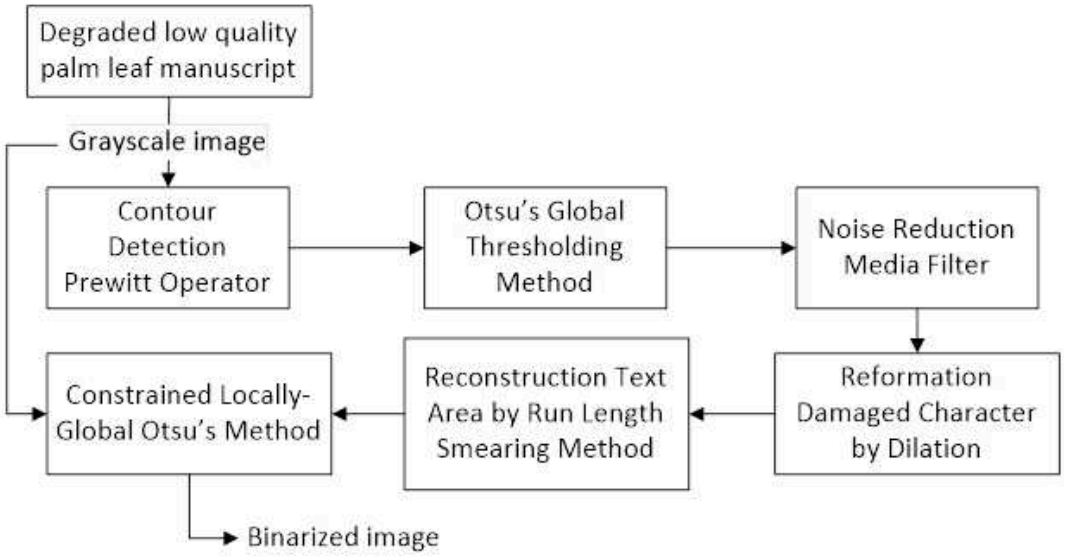


Figure 5.8: Semi-local binarization scheme [1]

Figure 5.9: Examples of extracted character area (on the left) and their semi-local binarization result (on the right) [1]

Figure 5.10: Original sample image, and sequence sample image of Prewitt, Otsu, Median Filter, Dilation, RLS Row, RLS Col, Local Otsu, Thinning, Pruning, Superposed Skeleton on Original Image [1]

ing method to the thinned characters image. A pruning method for the skeleton is effective to remove spurious unwanted parts of the skeleton, and it makes the manual correction process of the skeleton faster. Figure 5.10 shows a sample of an image sequence as the result of our specific scheme. The goodness of the results can only be estimated qualitatively by examining the results. Based on visual criteria, the proposed scheme provides a good initial image of skeleton with respect to image quality and preservation of meaningful textual character information.

Figure 5.11 shows the complete scheme of the proposed binarized images ground truth dataset construction. The detail quantitative evaluation including the analysis on binarized ground truth images variability will be reported in Sub Section 7.1.2.

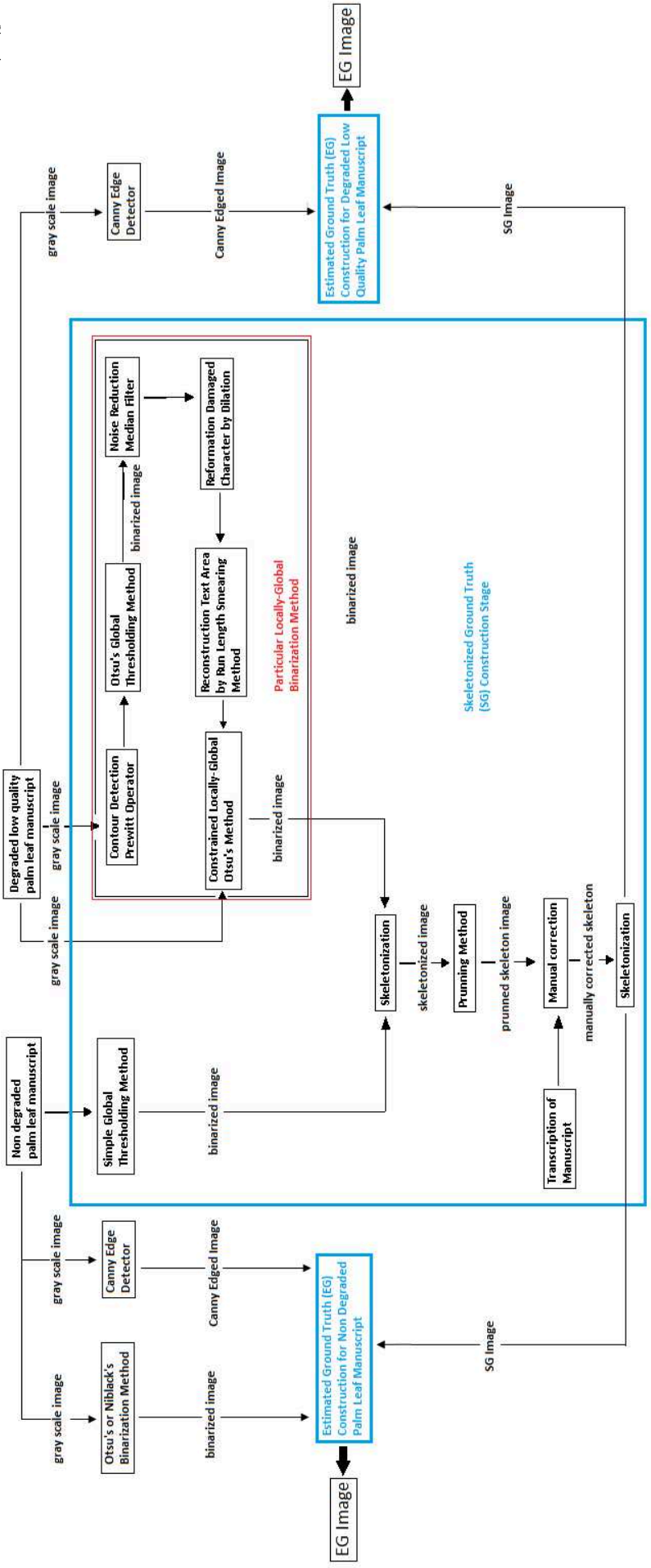


Figure 5.11: The complete scheme of the proposed binarized images ground truth dataset construction

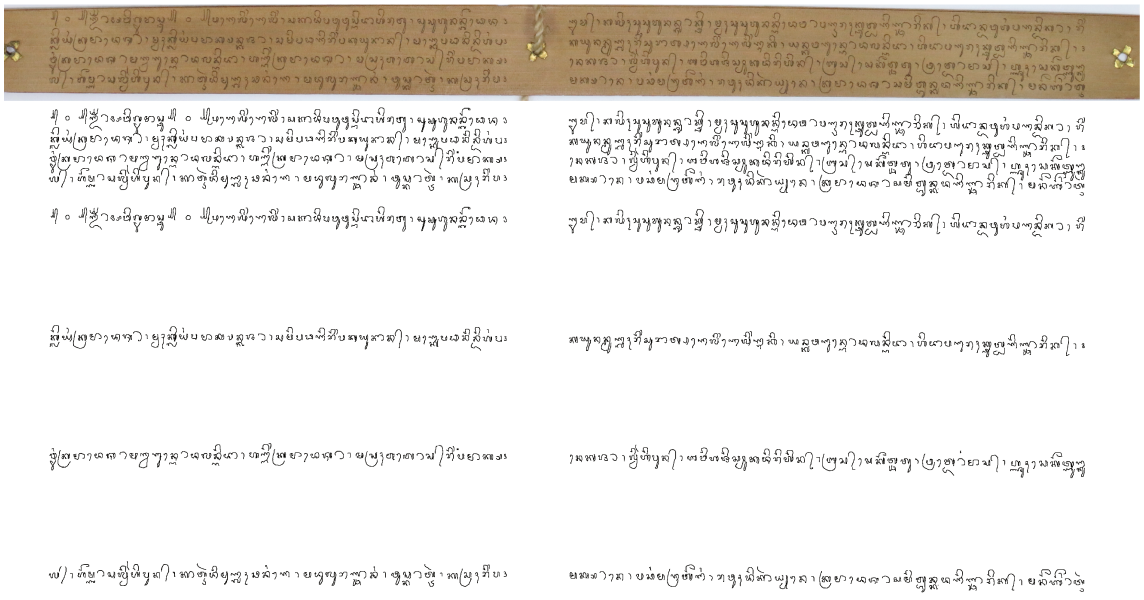


Figure 5.12: Text Line Segmented Images Ground Truth Sample (Original Image, Binarized Ground Truth Image, Text Line Image from 1st to 4th line)

5.3.3 Text Line Segmented Images Ground Truth Construction

The text line segmented images ground truth dataset (Figure 5.12) has been also generated manually by hand based on the binarized images ground truth dataset from the previous ground truthing process (see Sub Section 5.3.2).

5.3.4 Word Annotated Images Ground Truth Construction

To create the word annotated ground truth dataset of the manuscript, we used a collaboration work between the Balinese script philologists, students from the Department of Informatics, and students from the Department of Balinese Literature from Universitas Pendidikan Ganesha, Bali. The philologists read the manuscripts and created the Latin transliteration. Based on this Latin transliteration, a pair of students (one student in Informatics and one student in Balinese Literature) worked together to segment and to annotate each word in the manuscripts (Figure 5.13). The validation and correction of word annotations were done based on the expertise of the philologists. Any further verification remained open between the philologists and the ground truthers to correct and to validate the transcription during the annotation process.

We used Aletheia[2], an advanced document layout and text ground-truthing system [59], to segment and to annotate the words (Figure 5.14). After the segmentation and the annotation process, the manuscripts were then cropped based on word polygon coordinates in the XML file produced by Aletheia (Figure 5.15). To anticipate the limited size of digital image storage, we use the JPG format. For all word annotated images in this dataset, we used the filename format as follows (Figure 5.16):

word_filename_idword_cTL_rTL_cBR_rBR.jpg

where ***word*** indicated the word string of this word segment, ***filename*** indicated the orig-

[2]http://www.primaresearch.org/tools/Aletheia

Figure 5.13: Collaborative work between the Balinese script philologists, students from the Department of Informatics, and students from the Department of Balinese Literature to annotate the words

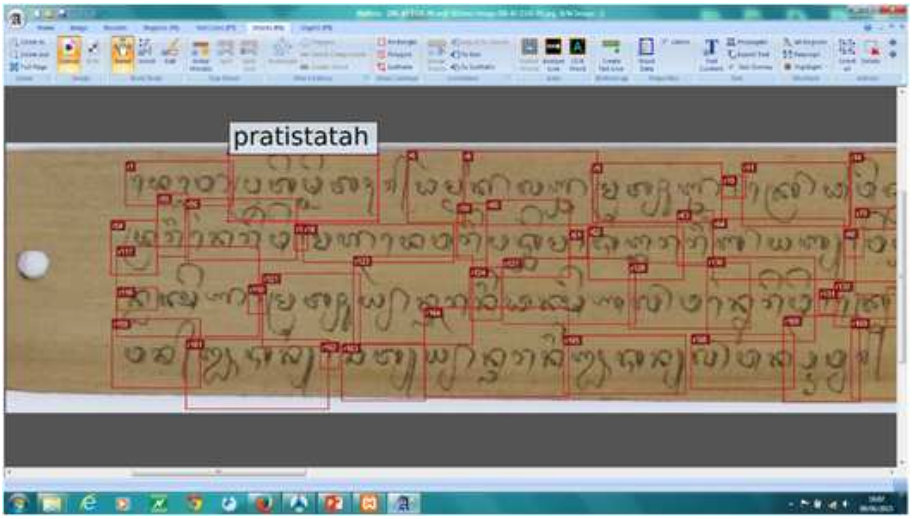


Figure 5.14: Word annotation with Aletheia [15]

inal manuscript of this word segment, ***idword*** indicated the ***id*** of this word segment (used only for the Aletheia software), ***cTL*** indicated the column coordinates of the top left point for this word segment, ***rTL*** indicated the row coordinates of the top left point for this word segment, ***cBR*** indicated the column coordinates of the bottom right point for this word segment, ***rBR*** indicated the row coordinates of the bottom right point for this word segment. The image coordinates column=1 and row=1 indicate the pixel on the top left corner of the image.

5.3.5 Isolated Glyph Annotated Images Ground Truth Construction

By using the collection of word annotated images which were produced in our previous ground truthing process (see Sub Section 5.3.4), we collected our isolated handwritten Balinese character dataset. First, we applied Otsu's [2, 3] binarization method to all word patch images. We automatically extracted all connected components found in the binarized word patch images. Our Balinese philologists then annotated manually all connected components that represent a correct character in Balinese script. To facilitate the work of the philologists, we developed a simple web based user interface for this character annotation process (Figure 5.17). With this web-based interface, more than one philologist can work together to verify, to correct and to validate the annotation of the characters. All annotated characters were displayed based on their given class. A hyperlink from each annotated character to their corresponding word annotated images

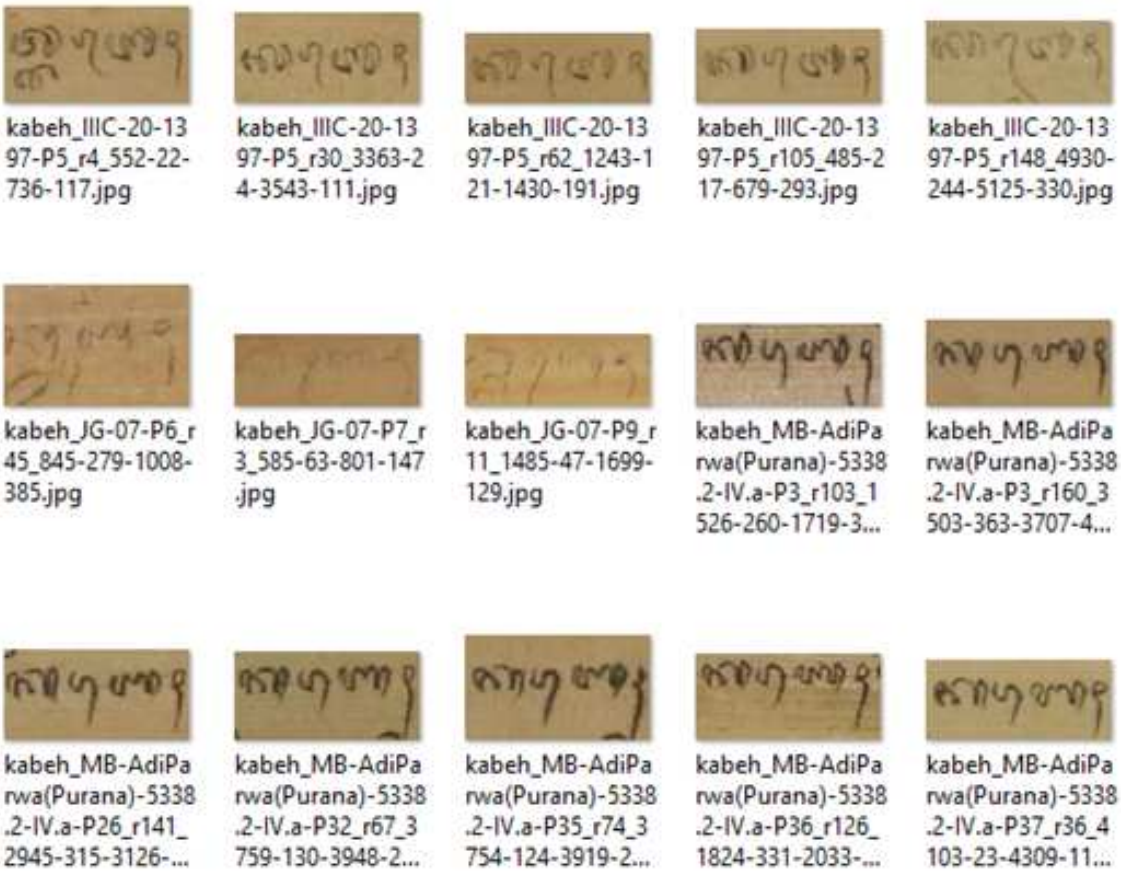


Figure 5.15: Samples of word annotated images [15]

WORD : endan

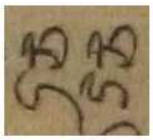
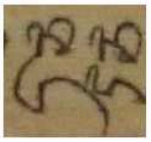
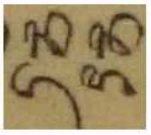
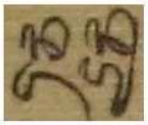
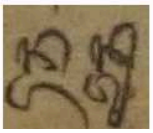
No	Filename	Image
1	endan_MB-AdiParwa(Purana)-5338.2-IV.a-P22_r104_3141-253-3257-357.jpg	
2	endan_MB-AdiParwa(Purana)-5338.2-IV.a-P31_r129_2145-389-2259-495.jpg	
3	endan_MB-AdiParwa(Purana)-5338.2-IV.a-P38_r67_3728-129-3842-229.jpg	
4	endan_MB-AdiParwa(Purana)-5338.2-IV.a-P39_r14_1821-29-1933-122.jpg	
5	endan_MB-AdiParwa(Purana)-5338.2-IV.a-P39_r115_4580-211-4697-313.jpg	

Figure 5.16: Word annotated images filename format

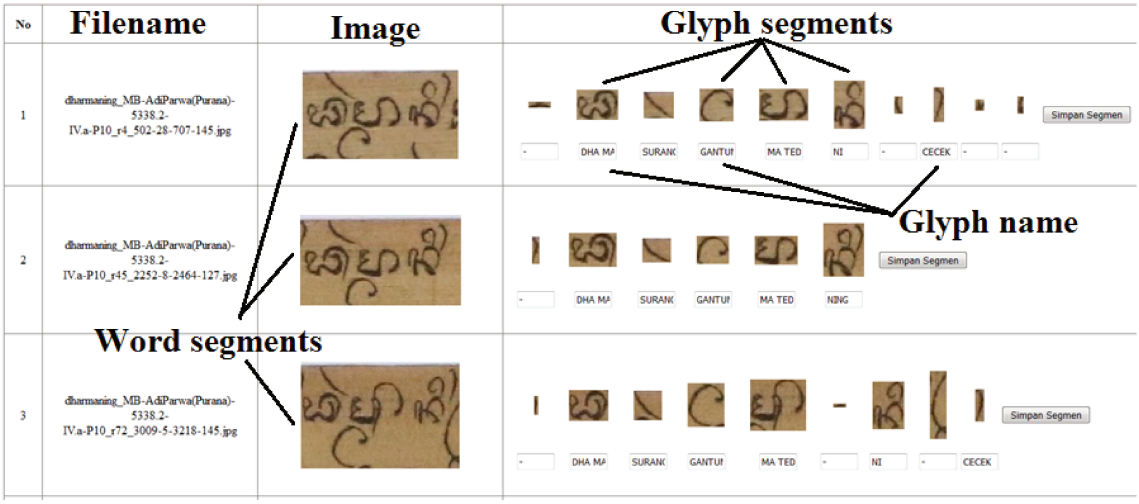


Figure 5.17: Screenshot of web based user interface for the character annotation process

was provided to allow the philologists to verify and to correct the annotation (Figure 5.18). Thumbnail samples of these character annotated images are shown in Figure 5.19. It should be noted that during this glyph annotation process, not only the basic glyphs (part of 156 complete glyphs set, see the description in Sub Section 2.3.4 and the complete list of these basic glyphs in Appendix C) were annotated, but some compound glyphs which were not separately segmented and were often found in the manuscript corpus were also annotated (see the list of these compound glyphs in Appendix E).

5.4 Dataset of AMADI_LontarSet

Under the scheme of the AMADI (Ancient Manuscripts Digitization and Indexation) Project, we have built the AMADI_LontarSet [15], the first handwritten Balinese palm leaf manuscript dataset. It includes eight components as follows: transliterated manuscript image ground truth dataset, binarized image ground truth dataset, text line segmented image ground truth dataset, word annotated image ground truth dataset, query-by-example word spotting dataset, isolated glyph annotated image ground truth dataset, query-by-example glyph spotting dataset and page image with glyph segmentation and recognition ground truth dataset. Six of those datasets (binarized images, text line segmented images, word annotated images, query-by-example word spotting, isolated glyph annotated images and query-by-example glyph spotting dataset) are already publicly available for scientific use at:

- http://amadi.univ-lr.fr/ICFHR2016_Contest/ [15]
- http://amadi.univ-lr.fr/ICDAR2017_Competition/
- http://amadi.univ-lr.fr/ICFHR2018_Contest/

Two other datasets will be published for the next competition.

390 pages of manuscripts from 23 different collections with 1266 total text lines are provided for the manuscript transliteration task.

Table 5.3 shows the summary of binarized images ground truth dataset for the AMADI_LontarSet [15]. Each image has two different ground truth binarized images. More detailed evaluation about the analysis of the ground truth binarized image variability of

Figure 5.18: Screenshot of character class verification

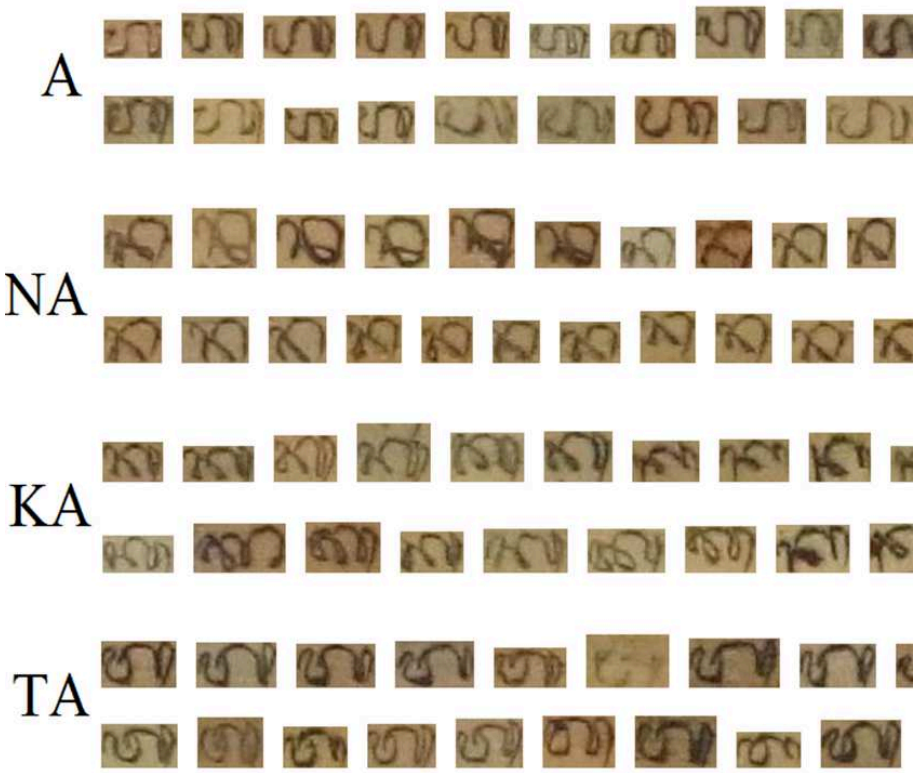


Figure 5.19: Samples of character-level annotated patch images of Balinese script on palm leaf manuscripts [15]

Table 5.3: Palm leaf manuscript dataset

Dataset	Train	Test	Description
Binarization	50 pages	50 pages	with the total of 2 x 100 binarized ground truth images
Text Line Segmentation Balinese 1	35 pages with 140 text lines	-	for non training based methods
Text Line Segmentation Balinese 2	Balinese 2.1: 47 pages with 181 text lines	Balinese 2.2: 49 pages with 182 text lines	for training based methods
Word recognition and transliteration	15,022 images from 130 pages	10,475 images from 100 pages	Text Latin
Isolated character/ glyph recognition	11,710 images	7,673 images	133 glyph classes

the palm leaf manuscripts will be given in Sub Section 7.1.2.3. The training set with the ground truth will be also provided for supervised learning methods.

The palm leaf manuscript datasets for text line segmentation task are also presented in Table 5.3. For training based methods, dataset Bali-2.1 can be used as the training set and dataset Bali-2.2 can be used as the testing set. The ground truth for line segmentation of the training set will be provided. The ground truth of each file is a raw image file. Each pixel stores a positive integer value corresponding to the ID of the text line it belongs to. For the background or undefined region, its pixel stores a zero value. The ground truth raw data format follows the evaluation tool of the ICDAR2013 Handwriting Segmentation Contest[3].

Table 5.3 shows also the summary of the word annotated image dataset for the AMADI_LontarSet [15] for the word recognition and transliteration task.

For the query-by-example word spotting task, the same dataset of 15,022 images from 130 pages of Word Annotated Images Ground Truth Dataset is provided for the training set. For the testing set, extracted from 100 other pages (different from the training set), 36 word annotated images are provided as a query test with their 257 word annotated images as query ground truth.

All patch images that have been segmented and annotated at the basic and compound glyph level constitute the isolated glyph dataset. Table 5.3 shows also the summary of isolated glyph annotated images dataset for the AMADI_LontarSet [15] for isolated character/glyph recognition task. The number of sample images for each class is different. Some classes are frequently found in our collection of palm leaf manuscripts, but some others are rarely used. As previously explained in Sub Section 5.3.5, the 133 glyph classes in this dataset consists of basic and compound glyph classes as follow.

- 74 basic glyph classes where 54 classes are listed in Unicode Table and 20 classes are not listed in Unicode Table. These 74 glyph classes are part of 156 complete glyphs set (see Sub Section 2.3.4 and the complete list in Appendix C).

- 59 compound glyph classes (see the complete list in Appendix E).

[3]http://users.iit.demokritos.gr/~nstam/ICDAR2013HandSegmCont/Protocol.html

Figure 5.20: Examples of Glyph spotting ground truth file

For query-by-example word spotting task, the dataset is extracted from the Isolated Glyph Annotated Image Ground Truth Dataset.

For the training set:

1. 10 original manuscript images with glyph character-level annotation (with the XML ground truth file format)
2. 19,383 glyph annotated images (without any information about the manuscript images) from 133 classes. It comes from the total glyphs for training and for testing in the Isolated Glyph Annotated Image Ground Truth Dataset.

For the testing set:

1. 10 original manuscript images
2. 250 glyph images as query test with the XML ground truth file

The XML ground-truth file for glyph annotated manuscript presented in Figure 5.20 is following the XML file format of the ICFHR 2016 Handwritten Keyword Spotting Competition (H-KWS 2016)[4]. A spotting area is defined by a rectangle with the TOP LEFT coordinate, the width and height.

19 manuscript pages with a complete glyph segmentation and annotation are provided to test and to evaluate the glyph segmentation and recognition scheme.

5.5 Additional Dataset of Khmer and Sundanese Palm Leaf Manuscripts

5.5.1 Corpus of Khmer and Sundanese Palm Leaf Manuscripts

In Cambodia, Khmer palm leaf manuscripts (Figure 5.21) are still seen in Buddhist establishments and are traditionally used by monks as reading scriptures. Various libraries and institutions have been collecting and digitizing these manuscripts and have even shared the digital images with the public. For instance, the École Française d'Extrême-Orient (EFEO) has launched an online database[5] [116] of microfilm images of hundreds

[4]https://www.prhlt.upv.es/contests/icfhr2016-kws/
[5]http://khmermanuscripts.efeo.fr

Figure 5.21: Khmer palm leaf manuscript

Figure 5.22: Sundanese palm leaf manuscript

of Khmer palm leaf manuscript collections. Some digitized collections are also obtained from the Buddhist Institute, which is one of the biggest institutes in Cambodia responsible for research on Cambodian literature and language related to Buddhism, and also from the National Library (situated in the capital city, Phnom Penh), which is home to a large collection of palm leaf manuscripts. Moreover, a standard digitization campaign was conducted in order to collect palm leaf manuscript images found in Buddhist temples in different locations throughout Cambodia: Phnom Penh, Kandal, and Siem Reap [117].

According to the era during which the documents were created, slightly different versions of Khmer characters are used in the writing of Khmer palm leaf manuscripts. The Khmer alphabet is famous for its numerous symbols (˜70), including consonants, different types of vowels, diacritics, and special characters. Certain symbols even have multiple shapes and forms depending on what other symbols are combined with them to create words. The languages written on palm leaf documents vary from Khmer, the official language of Cambodia, to Pali and Sanskrit, by which the modern Khmer language was considerably influenced. Only a minority of Cambodian people, such as philologists and Buddhist monks, are able to read and understand the latter languages.

The collection of Sundanese palm leaf manuscripts (Figure 5.22) comes from Situs Kabuyutan Ciburuy, Garut, West Java, Indonesia. The Kabuyutan Ciburuy is a complex of cultural heritage from Prabu Siliwangi and Prabu Kian Santang, the king and the son of the Padjadjaran kingdom. The cultural complex consists of six buildings. One of them is Bale Padaleuman, which is used to store the Sundanese palm leaf manuscripts. The oldest Sundanese palm leaf manuscript in Situs Kabuyutan Ciburuy came from the 15^{th} century. In Bale Padaleuman, there are 27 collections of Sundanese manuscripts. Each collection contains 15 to 30 pages, with dimensions of 25–45 cm in length x 10–15 cm in width [42].

The Sundanese palm leaf manuscripts were written in the ancient Sundanese language and script. The characters consist of numbers, vowels (such as a, i, u, e, and o), basic characters (such as *ha*, *na*, *ca*, *ra*, etc.), punctuation, diacritics (such as *panghulu*, *pangwisad*, *paneuleung*, *panyuku*, etc.), and many special compound characters.

5.5.2 Dataset

Table 5.4 shows the summary of binarized images ground truth dataset for Khmer and Sundanese palm leaf manuscripts. For Khmer manuscripts, one ground truth binarized image is provided for each image, but for the Sundanese manuscripts, each image has

Table 5.4: Palm leaf manuscript datasets for binarization task of Khmer and Sundanese palm leaf manuscripts

Manuscripts	Train	Test	Ground Truth
Khmer Extracted from EFEO [116, 118]	-	46 pages	1 x 46 pages
Sundanese Extracted from Sunda Dataset [42]	-	61 pages	2 x 61 pages

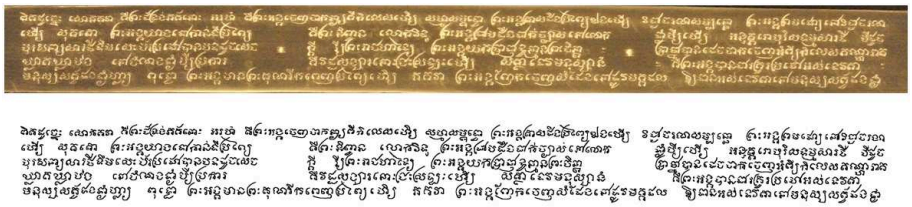


Figure 5.23: Khmer manuscript with binarized ground truth image

two different ground truth binarized images. In this research, we only used the first binarized ground truth image for evaluation. The binarized ground truth images for Khmer manuscripts were generated manually with the help of photo editing software (Figure 5.23). A pressure-sensitive tip stylus is used to trace each text stroke by keeping the original size of the stroke width [118]. The binarized ground truth images for Sundanese manuscripts were manually generated [42] using PixLabeler [17] (Figure 5.24). We used all images of the Khmer and Sundanese corpuses as a test set because the training-based binarization method (ICFHR Group 1 method, see Sub Section 4.3.1.3) was evaluated for the Khmer and Sundanese datasets by using only the pre-trained Balinese training set weighted model.

The Khmer and Sundanese palm leaf manuscript datasets for the text line segmentation task are presented in Table 5.5. The text line segmentation ground truth data for Sundanese manuscripts have been generated by hand based on the binarized ground truth images. For Khmer 1, a semi-automatic scheme is used [25, 118]. A set of medial points for each text is generated automatically on the binarization ground truth of the page image. Then those points can be moved up or down with a tool to fit the skew and fluctuation of the real text lines. We also note touching components spreading over multiple lines and the locations where they can be separated. For Khmer 2 and 3, an ID of the line it belongs to is associated with each annotated character. The region of a text line is the union of the areas of the polygon boundaries of all annotated characters composing it [117, 64].

Figure 5.24: Sundanese manuscript with binarized ground truth image

Table 5.5: Palm leaf manuscript datasets for text line segmentation task of Khmer and Sundanese palm leaf manuscripts

Manuscripts	Pages	Text Lines
Khmer 1 Extracted from EFEO [116, 25, 118]	43 pages	191 text lines
Khmer 2 Extracted from SleukRith Set [117, 64]	100 pages	476 text lines
Khmer 3 Extracted from SleukRith Set [117]	200 pages	971 text lines
Sundanese 1 Extracted from Sunda Dataset [25]	12 pages	46 text lines
Sundanese 2 Extracted from Sunda Dataset [42]	61 pages	242 text lines

Table 5.6: Palm leaf manuscript datasets for word recognition and transliteration task of Khmer and Sundanese palm leaf manuscripts

Manuscripts	Train	Test	Text
Khmer SleukRith Set [117]	16,333 images (part of 657 pages)	7,791 images (part of 657 pages)	Latin and Khmer
Sundanese Sunda Dataset [42]	1,427 images from 20 pages	318 images from 10 pages	Latin

Table 5.6 shows the summary of the word annotated image dataset for Khmer and Sundanese palm leaf manuscripts. For the Khmer dataset, all characters on the page have been annotated and grouped together into words (Figure 5.25). More than one label may be given to the created word. The order of how each character in the word is selected is also kept [117]. The Sundanese (Figure 5.26) word dataset was manually annotated using Aletheia [59].

Table 5.7 shows the summary of the isolated glyph annotated image dataset for the Khmer and Sundanese palm leaf manuscripts. The Sundanese character dataset was annotated manually [42] (Figure 5.27). For the Khmer character dataset, a tool has been developed to annotate characters/glyphs on the document page. The polygon boundary of each character is traced manually by dotting out its vertex one by one. A label is given to each annotated character after its boundary has been constructed [117] (Figure 5.28).

5.6 Conclusions

To develop the DIA system for palm leaf manuscripts, a corpus of palm leaf manuscripts and the construction of datasets was urgently needed. In this chapter, the proposed specific protocols to collect the palm leaf manuscript corpus and to construct the ground truth dataset were presented. These protocols were proposed to deal with the real situation and condition in Bali. They were designed to tackle to the difficulty to collect the manuscripts and to do the ground truthing process with the help of the philologists.

Figure 5.25: Khmer word dataset

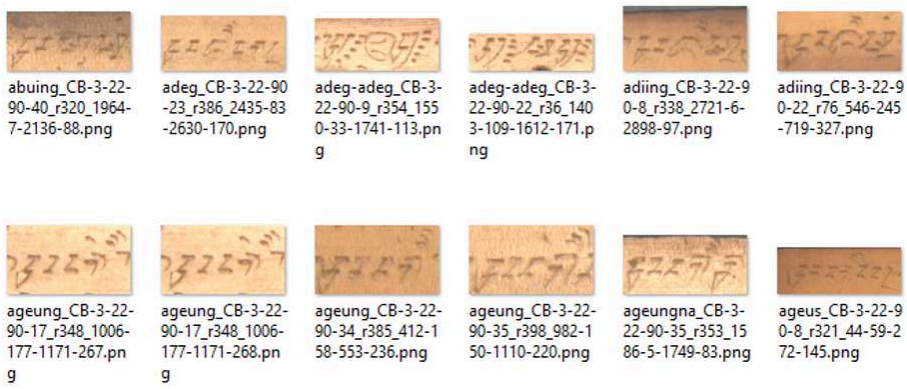


Figure 5.26: Sundanese word dataset

Table 5.7: Palm leaf manuscript datasets for isolated character/glyph recognition task of Khmer and Sundanese palm leaf manuscripts

Manuscripts	Classes	Train	Test
Khmer SleukRith Set [117]	111 classes	113,206 images	90,669 images
Sundanese Sunda Dataset [42]	60 classes	4,555 images	2,816 images

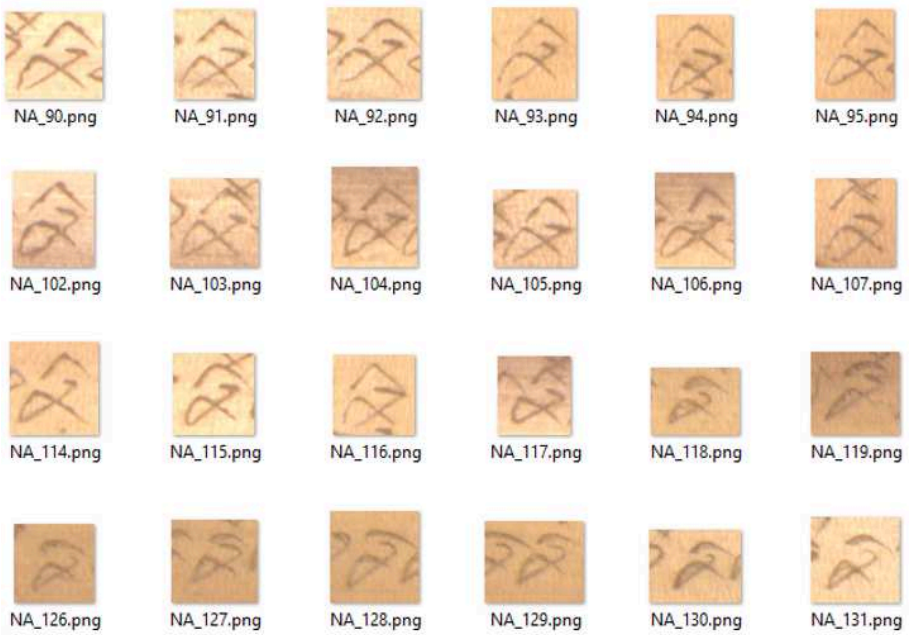


Figure 5.27: Sundanese character dataset

Figure 5.28: Khmer character dataset

Starting from the manuscript capturing process, some adaptions should be done to follow the socio-religious rules given by the museum or from the family as the owner of some private collections. For the ground truthing process of the manuscript corpus, some parts of the dataset were still constructed manually, while some other parts were created by the help of a semi-automatic process with some processing tools. During this ground truth dataset construction process, the collaboration between computer scientists and philologists took a very important role. As additional information, in the case of the Khmer and Sundanese corpus collections, the different conditions and difficulties could probably be found and different adaptation should be proposed and considered. In a near future, by using the first prototype of glyph recognition and transliteration that was developed in this work, for the second cycle after this PhD period we plan to develop the dataset in term of data quantity and variety to provide larger datasets for document analysis methods of palm leaf manuscripts.

Chapter 6

A Complete Scheme of Spatially Categorized Glyph Recognition for the Transliteration of Balinese Palm Leaf Manuscripts

This chapter presents the proposed complete scheme of spatially categorized glyph Recognition for the transliteration of Balinese palm leaf manuscripts. This chapter presents the detailed description of each of the six tasks of the proposed scheme, the design of a segmentation technique customized for this respective problem of manuscript analysis and the flow of recognition and the option selection rules for glyph recognition. This chapter also describes the knowledge representation and phonological rules which are built for the transliteration engine of Balinese script.

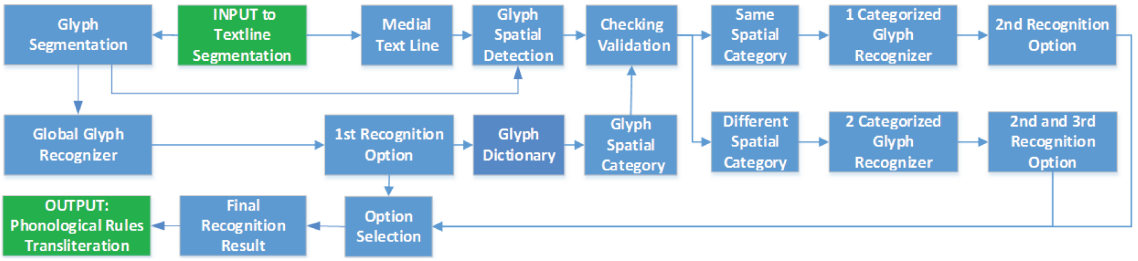


Figure 6.1: Proposed complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts

In the Khmer manuscript from Cambodia, in Balinese and in Sundanese manuscripts both from Indonesia, the writing system is based on a syllabic alphabet. With these scripts, some glyphs are written on the upper medial text line (as ASCENDER) or under the medial text line (as DESCENDER). The spatial position of each glyph relative to the medial text line on the manuscript can be adaptively used as an important information in glyph recognition. Taking into account the degraded condition of palm leaf manuscripts and the complexity of the Balinese script, in this chapter we present a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts. Generally, our proposed schema can be applied to these kinds of script. In this work, we focus on the glyph recognition and the transliteration of the Balinese script.

The sequence of analysis for glyph recognition and phonological rules for the transliteration of these manuscripts is presented on Figure 6.1. From the images of the manuscripts, this scheme is initialized with a text line and a glyph segmentation process. Then it is followed by a glyph recognition process. In this scheme, one global glyph recognizer and five different categorized glyph recognizers are combined in the checking and validation step to produce three glyph recognition options. These five different categorized glyph recognizers are based on the different spatial position of each glyph on the manuscript. These recognizers are used to verify and to validate the recognition result of the global glyph recognizer. Each glyph recognizer is built based on the combination of some feature extraction methods and it is trained on a single layer neural network. The trained network is initialized by an unsupervised feature learning process. The optional rules are applied to select the most appropriate glyph recognition option and are also used to detect the potential error in glyph segmentation and recognition. Finally, the selected glyph recognition option is sent to the phonological rules-based transliteration system to be transliterated to the Latin script.

The proposed complete scheme is a novelty in this work. The option selection rules to combine all the recognizers and the support to the transliteration engine are the main novelties.

6.1 Text Line Segmentation and Glyph Segmentation

First, the performance of six text line segmentation methods on the Southeast Asian palm leaf manuscript images are investigated [25] (the methods are described in Sub Section 4.3.2 and the results of evaluation will be reported in Section 7.2). Based on this comparative experimental studies, each method performed well on some specific characteristic of the manuscript collection. The behavior of some methods is greatly influenced by some

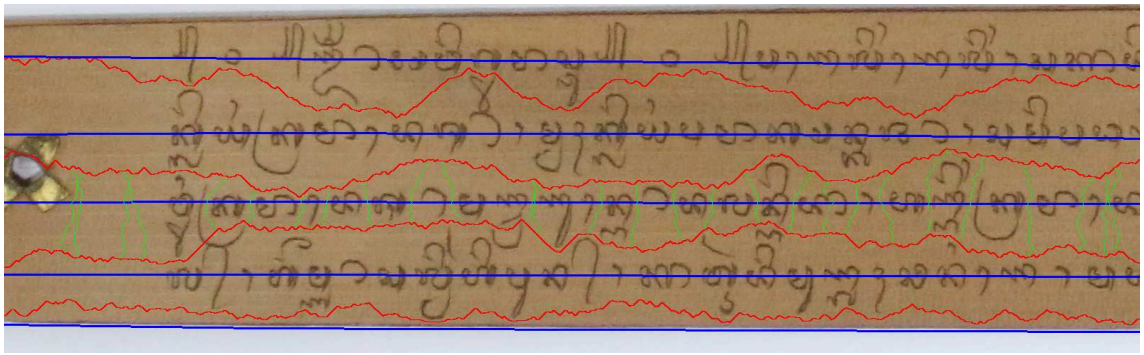


Figure 6.2: Text line segmentation and glyph area segmentation (green on the 3rd text line) with seam carving method. The medial text line (blue) and separating line (red) are both detected.

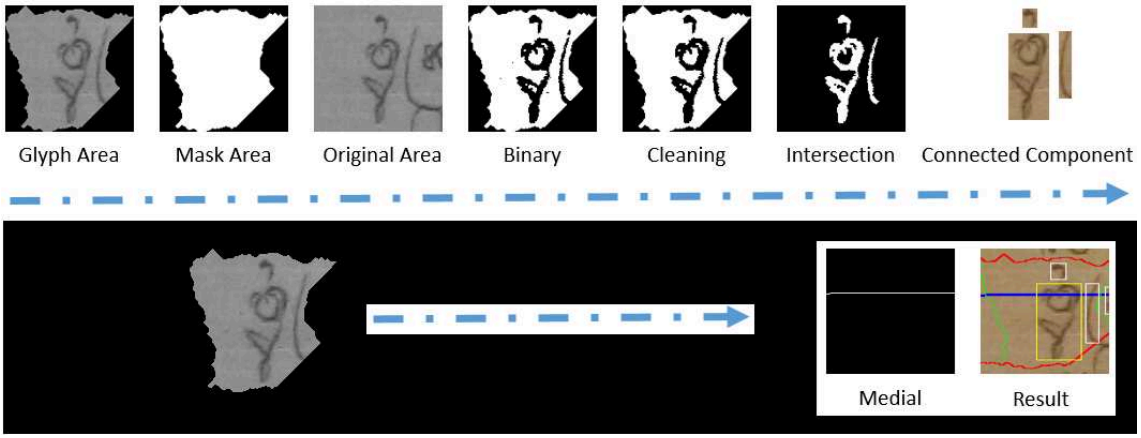


Figure 6.3: Glyph area detection and glyph segmentation process

challenges that are clearly present on each collection of the Southeast Asian manuscripts. For the Balinese manuscripts collection, the seam carving method [82, 83, 85] provides the best result on the grayscale text line segmentation [25]. In this method, two types of seams are calculated: the medial seams and separating seams (Figure 6.2).

In our schema, the seam carving method implementation [82] was not only applied to segment the text line, but also to segment the glyph area on each text line (Figure 6.2). The binarization with Otsu's method was then performed on this local area of the original grayscale manuscript. The Otsu's method seems more appropriate to determine the optimal threshold because the histogram in that local area is bimodal. The connected component analysis was finally used to clean up and to extract all glyph segments in each glyph area (Figure 6.3). A connected component is considered as a noise if the size (height or width) is less than a threshold defined as the minimum possible glyph size on the dataset of AMADI_LontarSet [15]. The glyph segmentation evaluation results will be presented in Table 7.14 in Section 7.4.

6.2 Detection of the Spatial Position for Glyph Category

For each detected glyph in a glyph area, their spatial position is defined relative to the medial text line on the manuscript. Six categories of spatial position for the glyph are defined (Figure 6.4). The algorithm is described in Algorithm 6.1. Ascender (***ASC***) glyphs

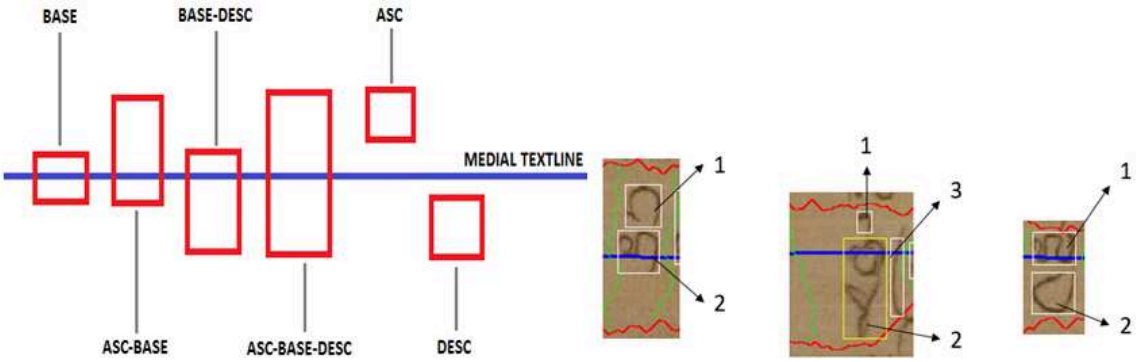


Figure 6.4: Left: Spatial position of the glyphs relative to the medial text line. Right: Examples of glyph ordering process

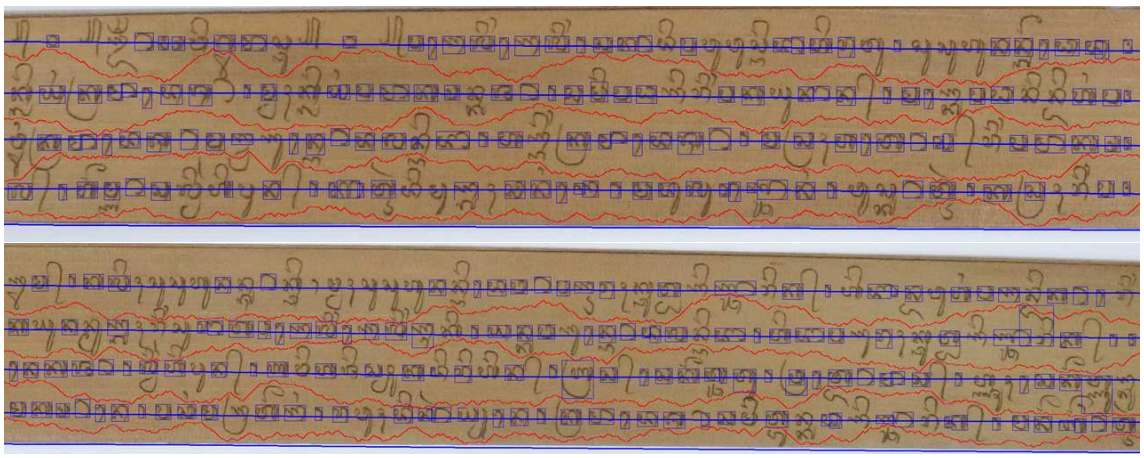


Figure 6.5: BASE Glyphs Detection

and descender (***DESC***) glyphs do not intersect the medial text line. ***ASC*** glyphs are written above the medial text line, and ***DESC*** glyphs are written below the medial text line. ***BASE*** glyphs are normally the basic glyph segments written exactly on the medial text line, while ***ASC-BASE*** glyphs and ***BASE-DESC*** glyphs are normally the compound glyph segments between a ***BASE*** glyph with an ***ASC*** glyph or with a ***DESC*** glyph. These spatial position categories are used to build the categorized glyph recognizer and to construct the phonological rules in the transliteration process. Figure 6.5-6.10 show the examples of spatial position for glyph category.

6.3 Glyph Ordering Process

After the glyphs were segmented in each glyph area, the glyph ordering process is then performed. The order of the glyphs is very important for the phonological rules-based machine to transliterate correctly the Balinese script. The glyphs which are on the medial text line (***BASE*** or ***ASC-BASE*** or ***BASE-DESC*** or ***ASC-BASE-DESC*** glyph) are ordered from left to right based on their left border position on the glyph area. If there is an ***ASC*** glyph and/or a ***DESC*** glyph, the ***ASC*** glyph will be placed before their associate ***BASE*** glyph, and the ***DESC*** glyph will be placed after their associate ***BASE*** glyph. This ordering rule is simply called as "***BASE-ASC-BASE-DESC-BASE Order***".

```
   Data: data input
   Result: data output
 1 if Intersection with MEDIAL TEXTLINE then
 2     There are 4 types : BASE, ASC-BASE, BASE-DESC, ASC-BASE-DESC. Check
        first ratio height and width
 3     if char_height ≥ 1.25 x char_width AND char_height ≥ 3 x min_h then
 4         char looks almost vertical and high enough: possibility of ASC-BASE,
            BASE-DESC, ASC-BASE-DESC
 5         if upper_part ≥ 1.5 x lower_part then
 6             level3name=ASC-BASE
 7         else
 8             if lower_part ≥ 1.5 x upper_part then
 9                 level3name=BASE-DESC
10             else
11                 Almost balance between two parts
12                 level3name=ASC-BASE-DESC
13             end
14         end
15     else
16         char looks almost square or horizontal (not too high)
17         level3name=BASE
18     end
19 else
20     (No-Intersection with MEDIAL TEXTLINE), There are 2 types : ASC, DESC
21     if Above MEDIAL TEXTLINE then
22         level3name=ASC
23     else
24         Below MEDIAL TEXTLINE
25         level3name=DESC
26     end
27 end
```

Algorithm 6.1: Detection of the Spatial Position for Glyph Category

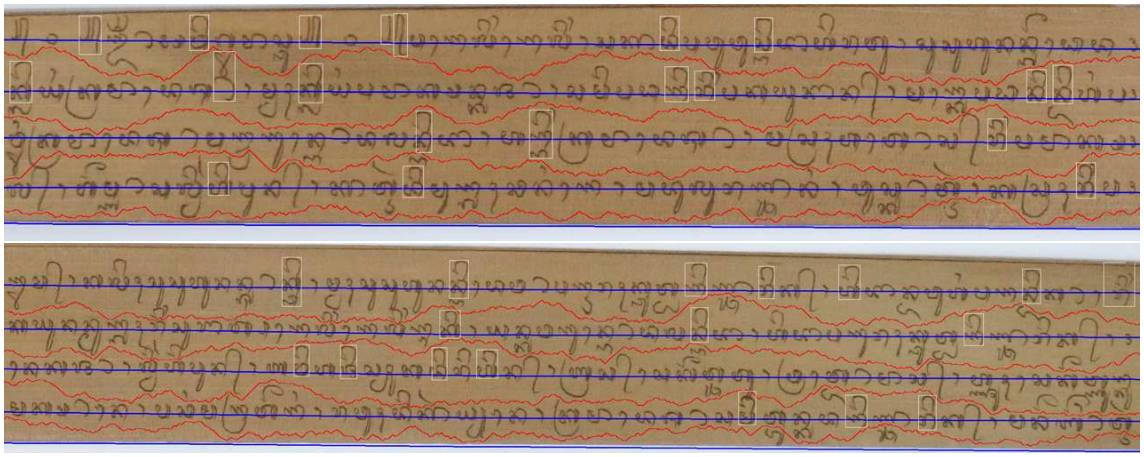


Figure 6.6: ASC-BASE Glyphs Detection

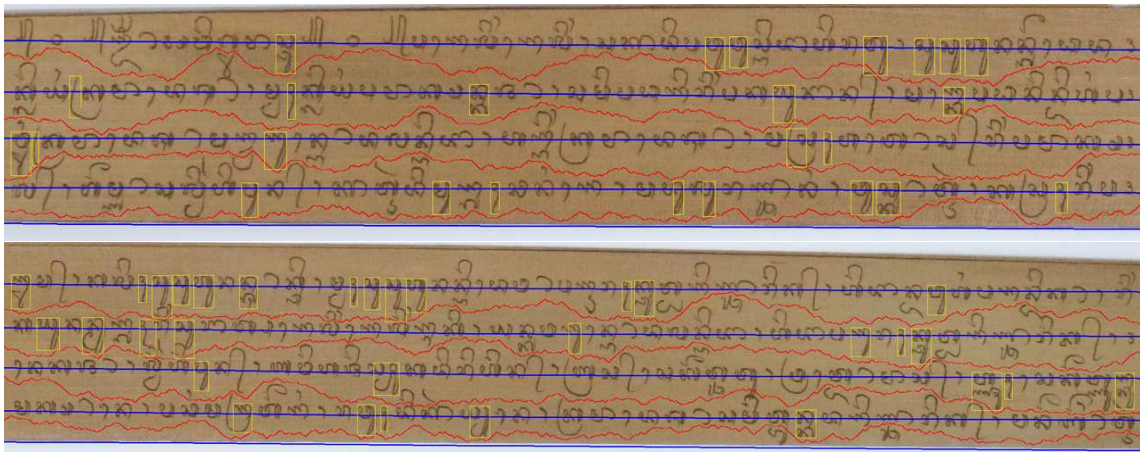


Figure 6.7: DESC-BASE Glyphs Detection

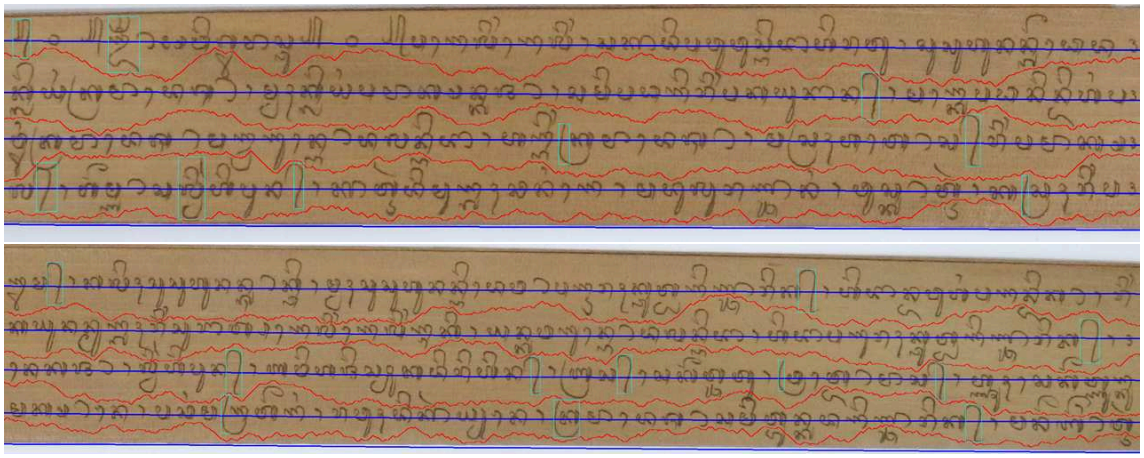


Figure 6.8: ASC-BASE-DESC Glyphs Detection

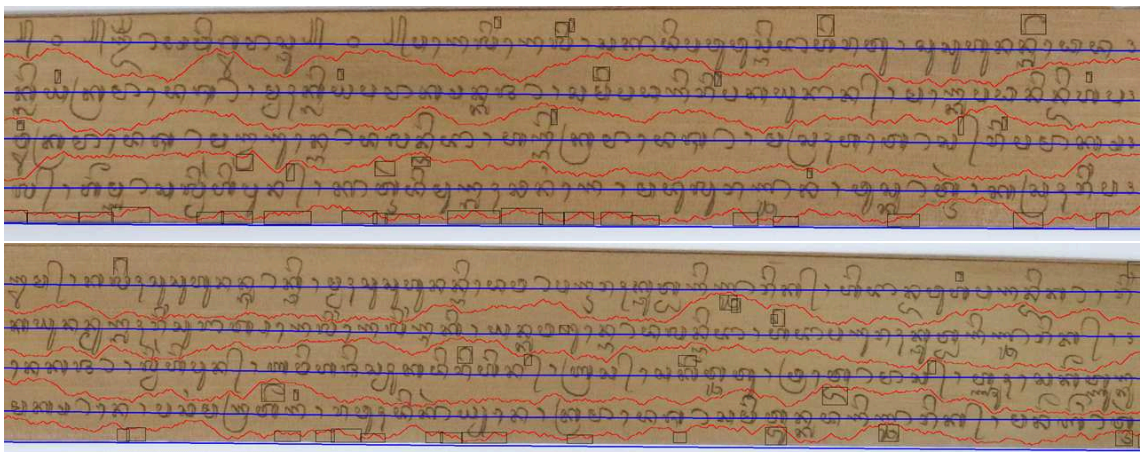


Figure 6.9: ASC Glyphs Detection

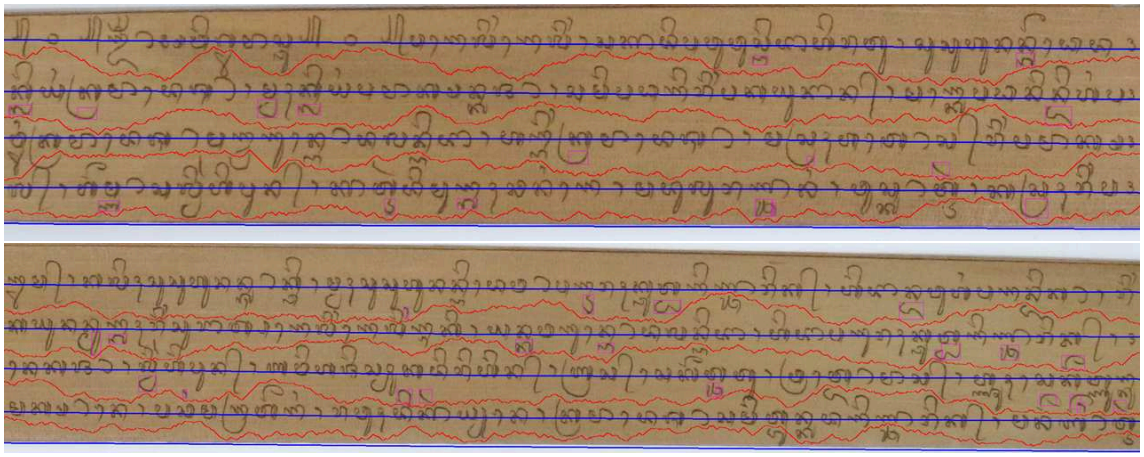


Figure 6.10: DESC Glyphs Detection

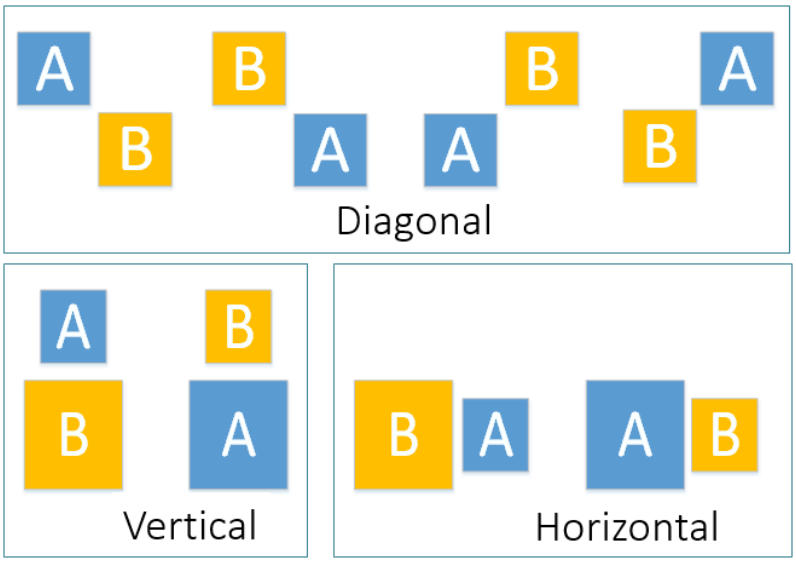


Figure 6.11: Spatial Relation between Glyphs

Because of the different handwriting styles, the position and the size of the ***ASC*** or ***DESC*** glyphs are sometimes not exactly on the vertical axis relatively to their associate ***BASE*** glyph position. It makes it difficult to define which ***BASE*** glyph belongs to which ***ASC*** or ***DESC*** glyph. To overcome this problem, the spatial relation between each pair of glyph is defined based on their four borders (top, bottom, left, and right) position. Two glyphs can be in a vertical relation, a horizontal relation, or a diagonal relation to each other (Figure 6.11). The nearest ***BASE*** glyph which can be associated to an ***ASC*** or a ***DESC*** glyph is then selected (Figure 6.12).

In Figure 6.4, on the right image for the second example, glyph 1 (***ASC***) and glyph 2 (***BASE-DESC***) are in vertical relation to each other, glyph 1 and glyph 3 (***BASE-DESC***) are in diagonal relation to each other, and glyph 2 and glyph 3 are in a horizontal relation to each other. Glyph 1 belongs to glyph 2, glyph 1 does not belong to glyph 3.

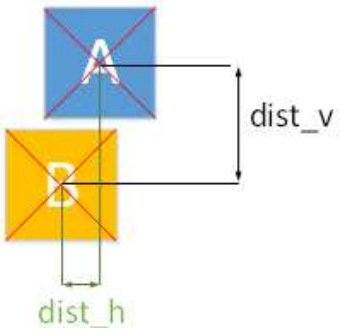


Figure 6.12: Distance Vertical and Horizontal between Glyphs

6.4 Glyph Recognition

For the glyph recognition task, we present our study and proposition on the isolated glyph recognition task, from the handcrafted feature extraction methods to the training based method with a neural network and the unsupervised feature learning, and the proposition on the global glyph recognition and categorized glyph recognition.

6.4.1 Proposed Combination of Features

The performance of the isolated glyph recognition greatly depends on the feature extraction step. To the best of our knowledge, only a few systems are available in the literature for Southeast Asian scripts recognition. With the aim of finding the combination of feature extraction methods for character recognition of Balinese script, we present our experimental study on feature extraction methods for character recognition on palm leaf manuscripts. We investigate and evaluate the performance of 10 feature extraction methods with two classifiers: k-NN (k-Nearest Neighbor) and SVM (Support Vector Machine) in 29 different schemes for Balinese script on palm leaf manuscripts [14]. The methods are described in Sub Section 4.3.3.1 and the experimental results will be reported in Table 7.10 in Section 7.3.

After evaluating the performance of those individual feature extraction methods, we found that the Histogram of Gradient (HoG) features as directional gradient based features [36, 89], the Neighborhood Pixels Weights (NPW) [87], the Kirsch Directional Edges [87] and Zoning method [39, 54, 87, 88] separately provide a very promising and good enough result. We obtained 84.35% of recognition rate by using only HoG features and 62.45% of recognition rate by using only Kirsch features. It means that the directional gradient based features and the four directional Kirsch edge images already serve as good feature discriminants for our dataset. The shape of Balinese characters is naturally composed by some curves. We can notice that the Kirsch edge image is able to give the initial directional curve features for each character. On the other hand, NPW features have an advantage that they can be applied directly to gray level images.

Trier et al. [92] reported that to improve the performance of an IHCR system, the combination of multi features is recommended. Our objective is to find the combination of feature extraction methods to adequately recognize the isolated characters of Balinese scripts on palm leaf manuscripts. Based on our preliminary experiment results, we proposed the proper and robust combination of feature extraction methods to increase the recognition rate. Our hypothesis is the four directional Kirsch edge images will provide

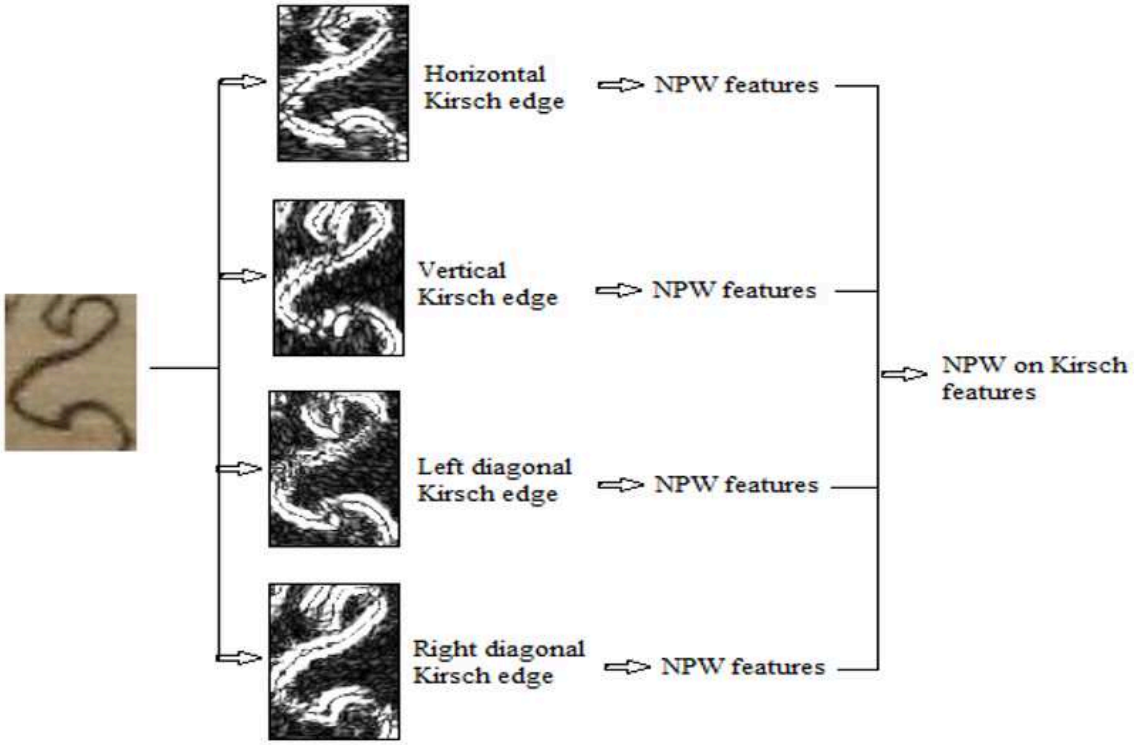


Figure 6.13: Scheme of NPW on Kirsch features [14]

a better feature discriminants for NPW features. Based on this hypothesis, we propose a new feature extraction method by applying NPW on kirsch edge images. We call this new method as NPW-Kirsch (Figure 6.13). Finally, we concatenate NPW-Kirsch with two other features, HoG and Zoning methods with k-NN as classifier. The study showed that the recognition rate can be significantly increased by applying NPW features on four directional Kirsch edge images. The use of NPW on Kirsch features in combination with HoG features and Zoning method can increase the recognition rate up to 85.16% with k-NN (nearest neighbours) classifier (see the experimental results in Table 7.10 in Section 7.3).

6.4.2 Training Based Method with Neural Network and Unsupervised Feature Learning

With the aim of improving the performance of our proposed feature extraction method, we continue our research investigation on isolated character recognition by implementing the neural network as classifier. In this second step, the same combination of feature extraction method was used and sent as the input feature vector to a single layer neural network character recognizer [16]. In addition to using only the neural network, we also applied an additional sub module for the initial unsupervised learning based on K-Means clustering (Figure 6.14). This schema was inspired by the study of Coates et al. [101, 102]. The unsupervised learning calculates the initial learning weight for the neural network training phase from the cluster centres of all feature vectors. The results of the experiment will be presented in Table 7.11 in Section 7.3.

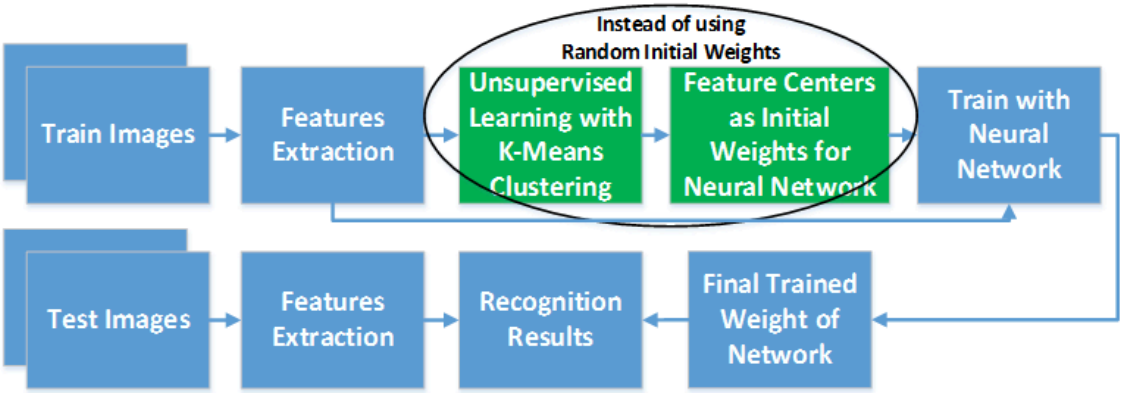


Figure 6.14: Schema of glyph recognizer with feature extraction method, unsupervised learning feature and neural network [16]

6.4.3 Global Glyph Recognition and Categorized Glyph Recognition

By using the same glyph recognizer schema (Figure 6.14), one global glyph recognizer and five different categorized glyph recognizers were built. The global recognizer was trained on a complete set of 133 glyph classes of Balinese script from AMADI_LontarSet [15]. The five different categorized glyph recognizers were trained only on a subset of glyph classes for each different spatial position category. The number of sample images for each class is different. Some classes are frequently found in our collection of palm leaf manuscripts, but some others are rarely used in Balinese language. The results of the experiment will be presented in Table 7.12 in Section 7.3.

6.5 Option Selection for Glyph Recognition

Based on the proposed scheme in Figure 6.1, the recognition schema for each segmented glyph is as follows:

- Send the glyph to the global recognizer (consider the result as the 1^{st} option).
- Based on this first recognition result, look up in the glyph dictionary (Appendix F) to find out to which spatial category this glyph should normally belong.
- Do the second recognition by sending this glyph segment to that categorized glyph recognizer (consider the result as the 2^{nd} option).
- Define the real spatial category of this glyph segment position relative to the medial text line. This real spatial category can be the same or be different with the spatial category verified in the glyph dictionary.
- Do the third recognition by sending this glyph segment to that categorized glyph recognizer (consider the result as the 3^{rd} option).
- In the end, three recognition options are then defined for each glyph: Global Recognition (***G***), Categorized Recognition based on Glyph Dictionary (***D***), and Categorized Recognition based on Glyph Spatial Position (***S***).

The option selection rules are as follow:

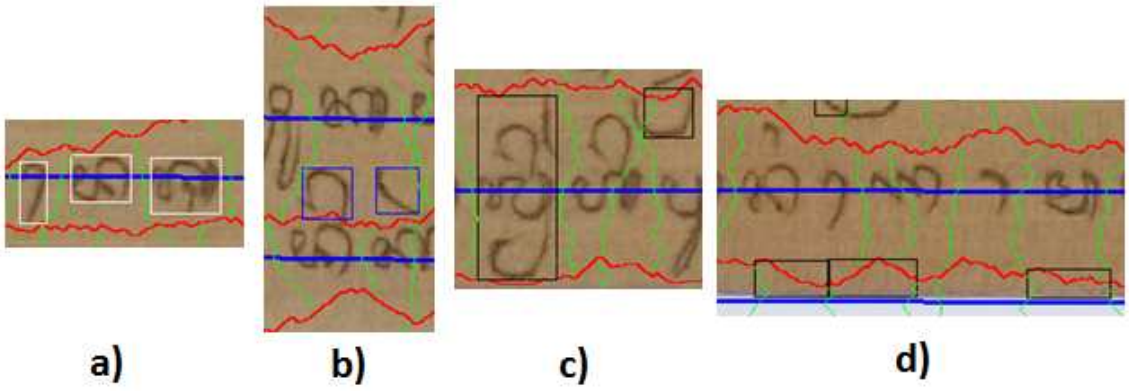


Figure 6.15: a) High confidence of correct segmentation recognition (white segments), b) Correct recognition but wrong line text assignment (blue segments), c) & d) Bad segmentation (black segments) [16]

- If the spatial category detection is the same with the glyph dictionary, there are two possibilities:
  - If $G = D = S$, there is only one option. It is a high confidence of correct segmentation and recognition. The final recognition result is $G/D/S$ (Figures 6.15a and 6.16).
  - If $G \neq (D = S)$, there are two different options. The recognition results are different between the global and the categorized recognizer. In this case, the final recognition result is D/S (Figure 6.17).
- If the spatial category detection is different than the glyph dictionary, there are three possibilities:
  - If $(G = S) \neq D$, there are two different options. This case is impossible, because to have G=S, G and D should have the same spatial category.
  - If $(G = D) \neq S$, there are two different options. There are three sub-cases. If (S = ***BASE***/***ASC-BASE***/***BASE-DESC***/***ASC-BASE-DESC*** and D = ***ASC***/***DESC***) or vice-versa, it means that there are a big difference between spatial category detection and the glyph dictionary. The final recognition result is S. If (S = ***ASC*** and D = ***DESC***) or vice-versa, it can be a correct recognition but a wrong text line assignment may be detected for this glyph segment (Figure 6.15b and 6.18). The final recognition result is G/D. For all other sub-cases, it may be a bad glyph segmentation (Figure 6.15c and 6.19). The final recognition result is G/D.
  - If $G \neq D \neq S$, there are three different options. This implies erroneous processing text line or glyph segmentation, so the spatial category detection is not correct (Figure 6.15d). The final recognition result is G. In the future, we could re-launch locally a specific treatment in this glyph segment.

Before applying the final option selection rules, two special cases were treated. The glyph "***CECEK***" and the glyph "***TALING***" can be written in two different spatial categories (Figure 6.20). For both glyphs, we only consider the spatial category from the glyph dictionary as follows:

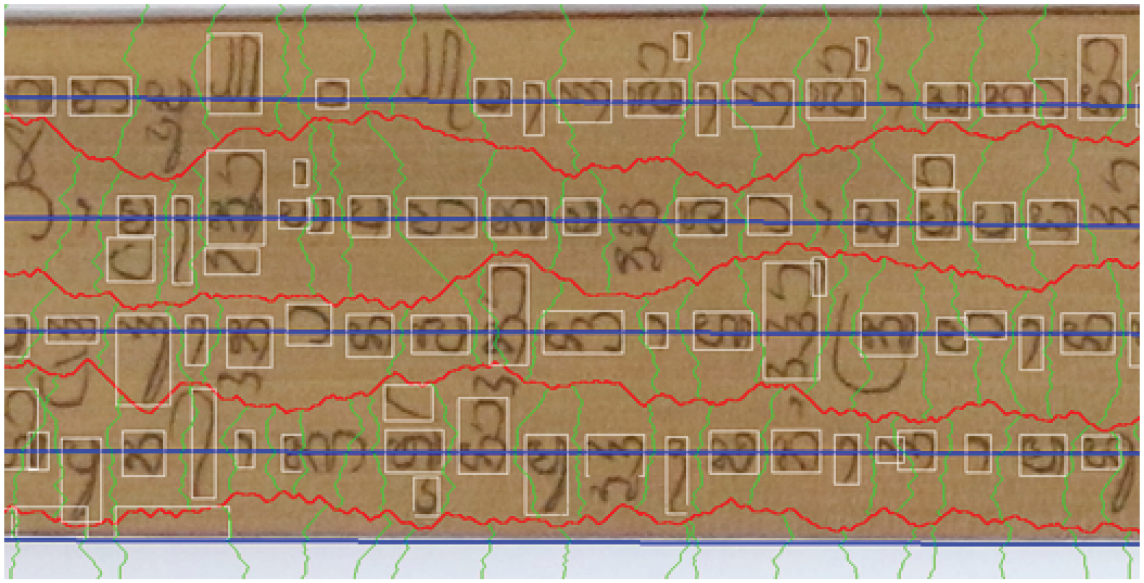


Figure 6.16: High confidence of correct segmentation and recognition

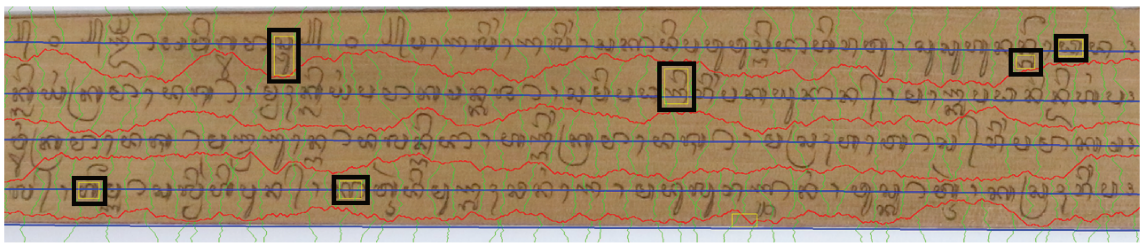


Figure 6.17: Confidence on recognition of D/S

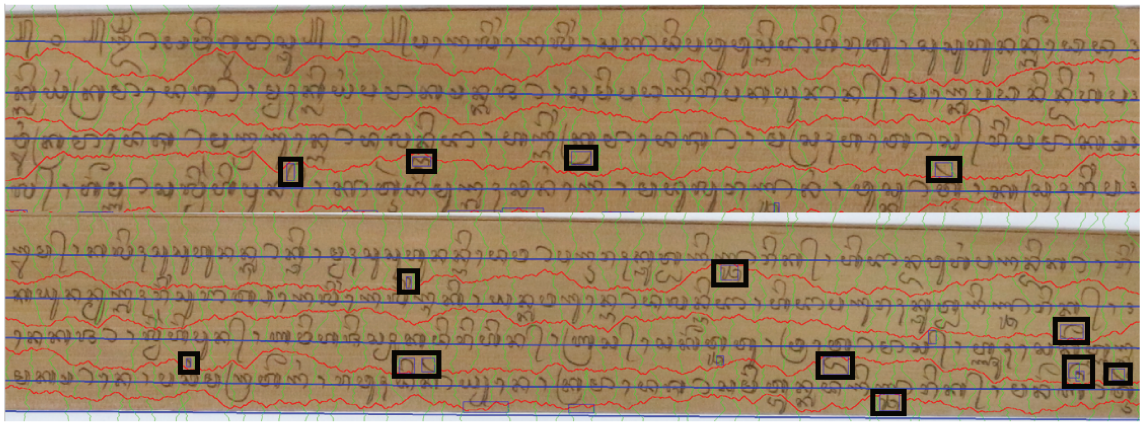


Figure 6.18: Confidence on recognition but potential error on text line segmentation

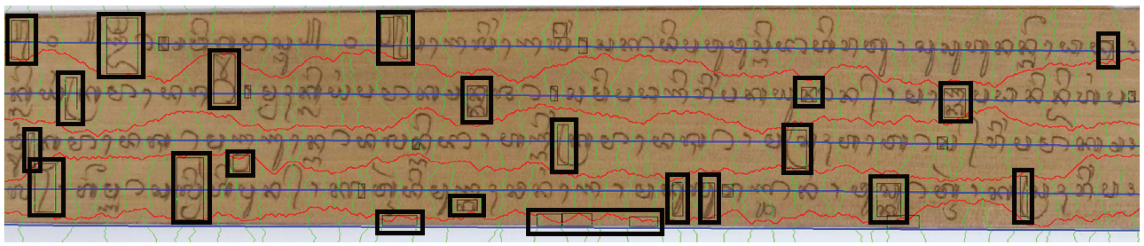


Figure 6.19: Bad segmentation, confidence on recognition of G/D

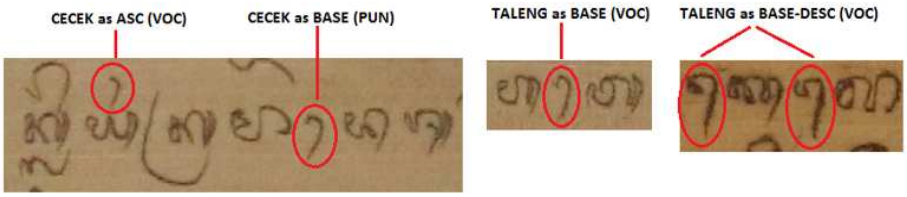


Figure 6.20: Special cases for glyph "***CECEK***" and glyph "***TALING***"

- Glyph ***CECEK*** can be written as ***ASC*** (***VOC***) or ***BASE*** (***PUN***) with exactly the same shape and size. But the dictionary only considers it as ***ASC*** (***VOC***). If ***CECEK*** is detected as ***BASE*** in ***S***, change ***S*** as ***D***.

- Glyph ***TALING*** can be written in two different sizes, a small size as ***BASE*** (***VOC***) and a bigger size as ***BASE-DESC*** (***VOC***). But the dictionary only considers it as ***BASE-DESC*** (***VOC***). If ***TALING*** is detected as ***BASE*** in ***S***, change ***S*** as ***D***.

6.6 Transliteration with Phonological Rule-based Machine

In this section, as a subset of DIA for Balinese palm leaf manuscripts, the knowledge representation and the phonological rules for the automatic transliteration will be described. A transliteration engine to transliterate the Balinese script into the Latin/Roman script is one of the most demanding systems which has to be developed for the collection of Balinese palm leaf manuscript images. In this section, we present an implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscript. In this system, a rule-based engine for performing the transliteration task is proposed. Our model is based on phonetics which are based on a traditional linguistic study of Balinese script transliteration. This automatic transliteration engine is needed to complete our proposed scheme of segmentation based transliteration with isolated glyph/character recognition (OCR) process on the palm leaf manuscript images. Sub Section 6.6.1 will present the detailed description of knowledge representation for the glyph segment image collection, the glyph properties and glyph categorizations. The phonological rules for Balinese script transliteration will be formally described in Sub Section 6.6.2. The results and evaluation of the experiments will be presented in Chapter 7 Sub Section 7.5.2.

6.6.1 Knowledge Representation

6.6.1.1 Glyph Segment Image Collection

The transliteration engine which is proposed in this work is based on and is developed for the isolated glyph annotated images collection of the AMADI_LontarSet [15] (see Section 5.4). From the 133 glyph classes in the collection which were used in the isolated Balinese glyph recognition task [15, 14], we group them in three categories of glyph as follow.

- Some of the glyphs represent the "basic" consonant of Balinese script. In their base form, they will produce a speech sound of a syllable which will be ended with a vowel "*A*". For example, glyph "*NA*", glyph "*CA*", glyph "*RA*", etc (Figure 6.21).

Figure 6.21: Consonant basic glyphs (from left to right: glyph "Na", "NA TEDONG", "GANTUNGAN NA", "TA", "TA TEDONG", and "GANTUNGAN TA")

  - Each of those consonant basic character glyphs has their associate second glyph form (conjunct form of consonants)[1] which is normally called *Gantungan* or *Gempelan*, or with another specific glyph name for *Pangangge Aksara*. For example, glyph "*GANTUNGAN NA*" for glyph "*NA*", glyph "*GUWUNG*" for glyph "*RA*", glyph "*GEMPELAN SA*" for glyph "*SA*". The *Gantungan* will be written under other consonant glyphs, and the *Gempelan* will be written aside other consonant glyph. The conjunct form of consonants will be used when one consonant follows another without a vowel in between[2]. These *Gantungan* and *Gempelan* will be used to annihilate the vowel "*A*" (the nucleus of the syllable) of their previous (upside or left side) written consonant glyph.
  - In some cases, it was found in the collection, the consonant of the basic character glyphs which were written in their cursive form (optional ligatures) with a glyph called "*TEDONG*". For example, glyph "*KA TEDONG*" for glyph "*KA*", glyph "*MA TEDONG*" for glyph "*MA*", glyph "*NA TEDONG*" for glyph "*NA*", etc. But the speech sound of the consonant will not be changed. Table 6.1 listed the consonant of the basic character glyphs, their associated second glyph forms (conjunct form), speech sound, and their component of syllable (onset and nucleus) which were found in the glyph segment image collection.

- In addition to the basic glyphs, in the glyph segment image collection, there are also the compound glyphs. A compound glyph is actually a glyph segment which is composed of more than one basic glyph, but the glyph segmentation process can not separate them (Figure 6.22). The collection of compound glyphs can facilitate the OCR process to deal with the improper segmentation task. The compound glyphs produce a speech sound of a syllable which will be ended with a different vowel than vowel "*A*" or will be combined with other consonant. Table 6.2 lists the consonant of the compound glyphs, their glyph components, and their component of syllables (onset, nucleus and coda) which were found in the glyph segment image collection.

- The rest of the glyphs consist of the numeral glyphs, punctuation glyphs, the vowel glyphs and special consonant glyph which can be used to change the speech of sound of other basic or compound glyphs (see the examples in Figure 6.23 and Table 6.3).

[1] In other words, each consonant has two forms, the regular and the appended form (https://www.loc.gov/catdir/cpso/romanization/balinese.pdf)

[2] www.omniglot.com/writing/balinese.htm

Table 6.1: Consonant basic glyphs and their second glyph form (conjunct form)

No	Consonant basic glyph name	Second glyph form name	Speech sound	Onset	Nucleus
1.	NA and NA TEDONG	GANTUNGAN NA	NA	N	A
2.	TA and TA TEDONG	GANTUNGAN TA	TA	T	A
3.	KA and KA TEDONG	GANTUNGAN KA	KA	K	A
4.	A and A TEDONG	GANTUNGAN A	A		
5.	WA and WA TEDONG	SUKU KEMBUNG	WA	W	A
6.	DA and DA TEDONG	GANTUNGAN DA	DA	D	A
7.	JA and JA TEDONG	GANTUNGAN JA	JA	J	A
8.	LA and LA TEDONG	GANTUNGAN LA	LA	L	A
9.	YA and YA TEDONG	NANIA	YA	Y	A
10.	MA and MA TEDONG	GANTUNGAN MA	MA	M	A
11.	SA		SA	S	A
12.	BA	GANTUNGAN BA	BA	B	A
13.	RA and RA TEDONG	GUWUNG	RA	R	A
14.	NGA	GANTUNGAN NGA	NGA	NG	A
15.	GA and GA TEDONG	GANTUNGAN GA	GA	G	A
16.	PA	GEMPELAN PA	PA	P	A
17.	DA MADU		DHA	DH	A
18.	CA	GANTUNGAN CA	CA	C	A
19.	NA RAMBAT and NA RAMBAT TEDONG		NA	N	A
20.	NYA	GANTUNGAN NYA	NYA	NY	A
21.	SA SAGA		SA	S	A
22.	SA SAPA		SA	S	A
23.	BA KEMBANG		BHA	BH	A
24.		GANTUNGAN TA LATIK	TA	T	A
25.	A KARA		A		
26.		GEMPELAN SA SAPA	SA	S	A
27.	I KARA		I		
28.	LA LENGA		LE		
29.	E KARA		E		
30.	TA TAWA	GANTUNGAN TA TAWA	TA	T	A
31.	U KARA		U		

Figure 6.22: Compound glyphs (from left to right: glyph “TU”, “KU”, “RU”, “DU”, “I”, “NI”, “TI”, “WI”)

Figure 6.23: Vowel glyphs (from left to right: glyph “TALING”, “TEDONG”, “ULU”, “SUKU”, “CECEK”, “PEPET”, “SURANG”)

Table 6.2: Consonant compound glyphs

No	Consonant compound glyph name	Glyph component	Speech sound	Onset	Nucleus	Coda
1.	TU	TA + SUKU	TU	T	U	-
2.	KU	KA + SUKU	KU	K	U	-
3.	I		I			-
4.	NI	NA + ULU	NI	N	I	-
5.	TI	TA + ULU	TI	T	I	-
6.	U		U			-
7.	RU	RA + SUKU	RU	R	U	-
8.	DU	DA + SUKU	DU	D	U	-
9.	WI	WA + ULU	WI	W	I	-
10.	DI	DA + ULU	DI	D	I	-
11.	WU	WA + SUKU	WU	W	U	-
12.	GU	GA + SUKU	GU	G	U	-
13.	NING	NA + ULU + CECEK	NING	N	I	NG
14.	KI	KA + ULU	KI	K	I	-
15.	LU	LA + SUKU	LU	L	U	-
16.	BU	BA + SUKU	BU	B	U	-
17.	YU	YA + SUKU	YU	Y	U	-
18.	TIA		TIA	T		-
19.	JU	JA + SUKU	JU	J	U	-
20.	NU	NA + SUKU	NU	N	U	-
21.	RI	RA + ULU	RI	R	I	-
22.	LI	LA + ULU	LI	L	I	-
23.	MU	MA + SUKU	MU	M	U	-
24.	SI	SA + ULU	SI	S	I	-
25.	PU	PA + SUKU	PU	P	U	-
26.	KNA	KA + GANTUNGAN NA	KNA	KN	A	-
27.	SU	SA + SUKU	SU	S	U	-

continued on next page

Table 6.2 – continued from previous page

No	Consonant compound glyph name	Glyph component	Speech sound	Onset	Nucleus	Coda
28.	GI	GA + ULU	GI	G	I	-
29.	NIA	NA + NANIA	NIA	N	IA	-
30.	NGU	NGA + U	NGU	NG	U	-
31.	KRA	KA + GUWUNG	KRA	KR	A	-
32.	JI	JA + ULU	JI	J	I	-
33.	JNA	JA + GANTUNGAN NA	JNA	JN	A	-
34.	BRA	BA + GUWUNG	BRA	BR	A	-
35.	MI	MA + ULU	MI	M	I	-
36.	GNA	GA + GANTUNGAN NA	GNA	GN	A	-
37.	WRE	WA + GUWUNG MACELEK	WRE	WR	E	-
38.	BI	BA + ULU	BI	B	I	-
39.	PI	PA + ULU	PI	P	I	-
40.	TRA	TA + GUWUNG	TRA	TR	A	-
41.	NGI	NGA + ULU	NGI	NG	I	-
42.	WUA	WA + SUKU KEMBUNG	WUA	W	UA	-
43.	CU	CA + SUKU	CU	C	U	-
44.	IA	A + NANIA	IA			-
45.	TUA	TA + SUKU KEMBUNG	TUA	T	UA	-
46.	GRA	GA + GUWUNG	GRA	GR	A	-

6.6.1.2 Glyph Properties and Categorizations

Seven properties to categorize the glyphs were specifically defined based on the existing glyph segment image collection.

- Property "***Id***" defined the identity number of the glyphs. The value of "***Id***" ranges from 1 to 133.
- Property "***Level1***" defined the name of the glyphs. The name of the glyphs normally represents their speech sound. But for some glyphs, the name is totally different.
- Property "***Level2***" is categorized in six groups:
  - ***CON*** for consonant,
  - ***VOC*** for vocal,
  - ***GAN*** for gantungan,
  - ***GEM*** for gempelan,
  - ***NUM*** for numeral, and
  - ***PUN*** for punctuation.

  GAN groups all second glyph form (conjunct form) of consonant basic glyphs (see Table 6.1) which should be written as a descender of other glyphs (below other

Table 6.3: Numeral, Punctuation, Vowel, and Special Consonant Glyphs

No	Glyph name	Type	Speech sound
1	0 - 9	Numeral	0-9
2	PAMADA	Punctuation	. (point)
3	TALING	Vowel	E
4	TEDONG	Vowel	Combined with TALING, it gives speech sound of "O"
5	ULU	Vowel	I
6	SUKU	Vowel	U
7	CECEK	Special Consonant and Punctuation	NG and , (comma)
8	PEPET	Vowel	E
9	SURANG	Special Consonant	R
10	ADEG-ADEG	Special Consonant	-
11	BISAH	Special Consonant	H
12	ULU SARI	Vowel	I
13	ULU CANDRA	Vowel	I
14	ULU RICEM	Special Consonant	AM
15	GUWUNG MACELEK	Special Consonant	RE
16	SUKU ILUT	Vowel	U

glyph). ***GEM*** groups all second glyph form (conjunct form) of consonant basic glyphs (see Table 6.1) which should be written aside (right side) of other glyph. A special consonant called "***ADEG-ADEG***" is also categorized as ***GEM*** because, if it is needed, it can also annihilate the vowel "***A***" (the nucleus of the syllable) of their previous glyph. ***VOC*** groups not only the vowel glyph on the Table 6.3. The special consonant glyphs were also categorized as ***VOC*** because they can change the speech sound of other consonant glyph. One special case is glyph "***CECEK***". It has two functions, as a special consonant and as a punctuation. In this system, glyph "***CECEK***" is considered as special consonant so it has the property of ***VOC***. All other basic and compound consonant glyphs are considered as ***CON***.

- Property "***Level3***" defined the spatial position of the glyphs related to the medial text line on the manuscript. It is used to construct the phonological rules for all glyphs from OCR results. Six categories of spatial position for the glyph were defined (Figure 6.24).

  - ***ASC*** (ascender) glyphs,
  - ***DESC*** (descender) glyphs,
  - ***BASE*** glyphs,
  - ***ASC-BASE*** glyphs,
  - ***BASE-DESC*** glyphs, and
  - ***ASC-BASE-DESC*** glyph.

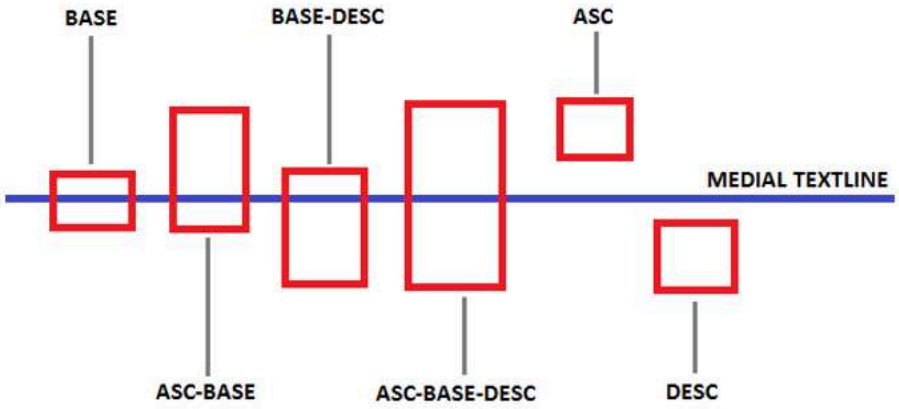


Figure 6.24: Spatial position of the glyphs related to the medial text line

ASC (ascender) glyphs and ***DESC*** (descender) glyphs do not intersect the medial text line, ***ASC*** glyphs are written above the medial text line, and ***DESC*** glyphs are written below the medial text line. ***BASE*** glyphs are normally the basic glyph segments written exactly on the medial text line, while ***ASC-BASE*** glyphs and ***BASE-DESC*** glyphs are normally the compound glyph segments between a ***BASE*** glyph with an ***ASC*** glyph or with a ***DESC*** glyph. In the glyph collection, there is only one ***ASC-BASE-DESC*** glyph class which is called glyph "***ADEG-ADEG***".

- Property "***StartSyllable***" keeps the information of the onset of the syllable for the consonant basic glyphs or the speech sound for the consonant compound glyphs, numeral, punctuation, and special consonant glyphs.
- Property "***EndSyllable***" keeps the information of the nucleus of the syllable for the consonant basic glyphs (see Table 6.1).
- Property "***SplitSyllable***" keeps the information of the onset, nucleus, and coda of the syllable for the consonant compound glyphs (see Table 6.2). All information about glyph properties and categorizations are saved in a XML file and will be loaded on a list of glyphs with a linked list pointer data structure.

The glyph properties and categorizations are stored in glyph dictionary (see Appendix F).

6.6.2 Phonological Rules

The method of transliteration depends on the characteristics of the source and target scripts [108]. Our model is based on phonetics which are based on traditional linguistic study of Balinese transliteration. Based on the glyph segment image collection, the glyph properties and categorizations, the conditional scheme of phonological rules for the transliteration of Balinese script have been finally identified and formally defined. The phonological rules defining the transliteration are provided and are applied in sequential conditional checking order. The engine checks all conditional phonological rules that might be applied to the given sequential glyph position and apply them in sequential order one by one. The OCR module for Balinese glyph recognition feeds the transliteration module with a sequential glyph data structure which is presented in Figure 6.25. In Balinese script, the final speech sound for a syllable of a current (***CURR***) base (***BASE***) glyph will be determined by :

- the ascender (***ASC***) of current glyph,
- the descender (***DESC***) of current glyph,
- the ***BASE*** of the ***NEXT*** glyph,
- the ***BASE*** of the previous (***PREV***) glyph,
- or even in some certain phonological rules, it can also be influenced by the ***BASE*** of the two previous (***PREV2***) glyphs.

For example a rule for ***TALING*** and ***TEDONG*** :

- *The **CONSONANT** can apply the rule of **TALING** and/or **TEDONG** if its next is not a **GEMPELAN**. If its next is a **GEMPELAN**, than the rule of **TALING** and/or **TEDONG** will be applied for that next **GEMPELAN**. Meanwhile, the **GEMPELAN** can take into account the rule of **TALING** and/or **TEDONG** if and only if **TALING** can be found in the two previous position of this **GEMPELAN***

The phonological rules for Balinese script transliteration should be finally built and be formally defined based on that OCR output data structure as follows. (Only a few examples of the phonological rules are described in this sub section. The complete rules can be found in Appendix G).

Examples of the phonological rules for transliteration of Balinese script:

- **RULE1**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE ⇒ SPEECH_SOUND = CURR.BASE.STARTSYLLABLE
- **RULE2**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = SPEECH_SOUND + CURR.DESC.STARTSYLLABLE
- **RULE3**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL2 ≠ GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ”TEDONG” ⇒ SPEECH_SOUND = SPEECH_SOUND + PREV.BASE.STARTSYLLABLE
- **RULE4**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL2 ≠ GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 = ”TEDONG” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”O”
- **RULE5**: IF PREV2.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 ≠ CON AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ”TEDONG” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”E”
- **RULE6**: IF PREV2.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 ≠ CON AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 = ”TEDONG” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”O”
- **RULE7**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ”NANIA” ⇒ SPEECH_SOUND = SPEECH_SOUND + NEXT.BASE.STARTSYLLABLE
- **RULE8**: IF PREV.BASE.LEVEL1 ≠ ”TALENG” AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = BASE AND CURR.BASE.ENDSYLLABLE = ”A” AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ”ADEG-ADEG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”A”

For the example in Figure 6.25, when the transliteration engine precedes the gylph “***MA***” as the ***CURR BASE*** glyph, ***RULE1*** will be applied first, because glyph “***MA***” is a ***CON BASE*** type glyph. In this case, the transliteration engine will take first the ***START-SYLLABLE*** of glyph “***MA***” as the speech sound output, which is “***M***”. The second rule which can be applied to this sequential glyph position for glyph “***MA***” is ***RULE8***. ***RULE8***

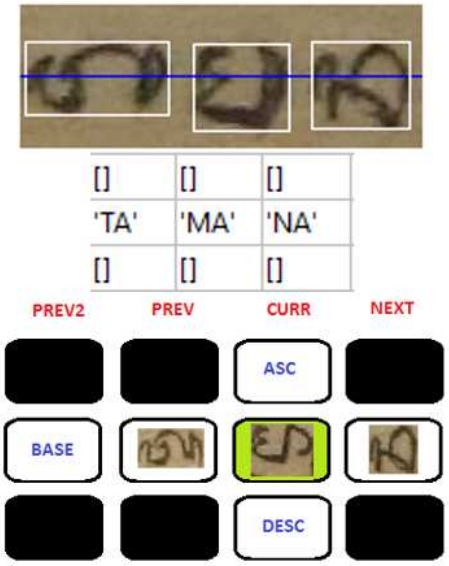


Figure 6.25: Example of RULE1 and RULE8 which are applied to an OCR result

describes the concatenation of speech sound "***A***" with the last speech sound output, because the ***CURR BASE*** glyph does not have an ***ASC*** glyph and/or a ***DESC*** glyph. The ***PREV BASE*** glyph is not a glyph "***TALING***" and the ***NEXT BASE*** glyph is not a glyph "***ADEG-ADEG***" or ***GEM*** type glyph. The final speech sound output will be "***MA***". If all the same sequential rules can be applied to glyph "***TA***" on the left and glyph "***NA***" on the right of glyph "***MA***", the final transliteration output will be "***TAMANA***". The experimental results of this phonological rule based transliteration will be presented in Table 7.15 in Sub Section 7.5.2.1.

The glyph ordering, recognition and options selection module finally feeds the transliteration machine with a sequential glyph data structure which is presented in Figure 6.25.The phonological rules defining the transliteration are finally applied in sequential conditional checking order. The engine checks all conditional phonological rules that might be applied to the given sequential glyph position and apply them in sequential order one by one. The experimental results of segmentation based transliteration are presented in Table 7.16 in Sub Section 7.5.2.

6.7 Conclusions

A complete scheme for glyph recognition and phonological rules for the transliteration of these manuscripts are presented in this work. The proposed scheme consists of six tasks: the text line and glyph segmentation, the detection of the spatial position for glyph category, the glyph ordering process, the global and the categorized glyph recognition, the option selection for the glyph recognition and the transliteration with phonological rules-based machine. For this scheme, five different categories of glyph recognizers based on the spatial positions on the manuscript are proposed. These recognizers are used to verify and to validate the recognition result of the global glyph recognizer. Each glyph recognizer is built based on the combination of some feature extraction methods and it is trained on a single layer neural network. The trained network is initialized by an unsupervised feature learning. The output of the glyph recognition scheme will be sent as the input to the phonological transliteration system.

Chapter 7

Experiments and Discussion

This chapter summarizes all experimental evaluations which have been done in this research work and the discussions of the results for each task and step in the DIA system for Balinese palm leaf manuscripts. In this chapter, the performance of each method for the DIA tasks on palm leaf manuscript collections are presented.

7.1 Binarization Evaluation

7.1.1 Evaluation Metrics

Three metrics of binarization evaluation proposed in the DIBCO 2009 contest [114] are used in this analysis to measure the difference between evaluated binarized image and ground truth binarized image, or between two ground truth binarized images from two different ground truthers. Those three metrics are F-Measure (FM), Peak SNR (PSNR), and Negative Rate Metric (NRM) [18].

F-Measure (FM)
FM is defined from Recall and Precision,

$$Recall = \frac{TP}{FN + TP} * 100 \tag{7.1}$$

$$Precision = \frac{TP}{FP + TP} * 100. \tag{7.2}$$

TP is defined as true positive, occurs when the image pixel is labeled as foreground and the ground truth is also. FP is defined as false positive, occurs when the image pixel is labeled as foreground but the ground truth is labeled as background. FN is defined as false negative, which occurs when the image pixel is labeled as background but the ground truth is labeled as foreground. FM is defined as

$$FM = \frac{2 * Recall * Precision}{Recall + Precision}. \tag{7.3}$$

A higher F-measure indicates a better match.

Peak SNR (PSNR)
PSNR is calculated from Mean Square Eror (MSE),

$$MSE = \sum_{x=1}^{M} \sum_{y=1}^{N} \frac{(I_1(x,y) - I_2(x,y))^2}{M * N} \tag{7.4}$$

$$PSNR = 10 * \log_{10}(\frac{C^2}{MSE}) \tag{7.5}$$

where C is defined as 1, the difference between foreground and background colors in the case of binary image. A higher PSNR indicates a better match.

Negative Rate Metric (NRM)
NRM is defined from the negative rate of false negative (NR_{FN}) and negative rate of false positive (NR_{FP}),

$$NR_{FN} = \frac{FN}{FN + TP} \tag{7.6}$$

$$NR_{FP} = \frac{FP}{FP + TN}. \tag{7.7}$$

TN is defined as true negative, which occurs when both the image pixel and ground truth are labeled as background. TP, FN, and FP are the same definitions as in F-Measure.

NRM is defined as

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \tag{7.8}$$

The value of NRM when we assumed the image drawn by the first ground truther is ground truth image will not be the same with the value of NRM when we assumed the image drawn by the second ground truther is ground truth image. In this case, we calculated two value of NRM: NRM1 and NRM2. A lower NRM indicates a better match.

7.1.2 Binarized Image Ground Truth Dataset Evaluation: Results and Discussion

7.1.2.1 Experiments for Nondegraded and Degraded Low Quality Palm Leaf Manuscripts

We experimentally tested the framework for construction of ground truth binarized image for nondegraded and degraded low quality palm leaf manuscripts [1]. For this initial experimental study, we only used the available sample scanned images from the Museum Bali, Museum Gedong Kirtya, and from private family collections. The sample images for the nondegraded palm leaf manuscripts come from the story of Kakawin Ramayana (reference code 07.41/5634.3 and 07.40/5634.3), in the Museum Bali, Denpasar, Indonesia. The palm leaf manuscripts were ordinarily scanned with scanner Epson L210 in 600 dpi without any specific external illumination. The sample images for the degraded low quality palm leaf manuscripts come from the private family collection and Museum Gedong Kirtya, Singaraja, Indonesia. The manuscripts were scanned at 300 dpi and it consists of several types of degradations as already mentioned in the previous chapter. The manuscripts were written on both sides, but there was no back-to-front interference observed.

For nondegraded palm leaf manuscripts, we used the simplest and the most conventional global thresholding method with a proper threshold selected manually to obtain the initial binarized image. With this initial binarized image, it is already sufficient to obtain an acceptable skeletonized image. We performed the manual correction of the skeleton, guided by the transcription of the manuscript provided by a philologist, to finally obtain the skeleton ground truth of the manuscript. Figure 7.1 shows a snapshot of a simple prototype with user friendly interface that we developed and used to facilitate the manual correction process. We finally constructed the ground truth image by dilating the corrected skeleton image, constrained by the Canny edge image and the initial binarized image from Otsu's global method. We use Otsu's global method instead of the same global fixed thresholding method used in our skeleton ground truth construction because we need a more complete connected component of all characters detected on the binarized image. Other binarization methods can also be used, for example, Niblack's method or the multi resolution version of Otsu's method [6]. They also provide a satisfactory preliminary binarized image. Figure 7.2 shows an example of final ground truth image from a nondegraded palm leaf manuscript. It is visually an acceptable estimated ground truth image for the manuscript.

For degraded low quality palm leaf manuscripts, we applied our proposed specific binarization scheme (see Sub Section 5.3.2.1) by defining the optimal value of parameters based on our empirical experiments as follow: filter size 3x3 for Median Filter, square structuring element size 3x3 for Dilation, smearing 3 pixels in row and 3 pixels in column

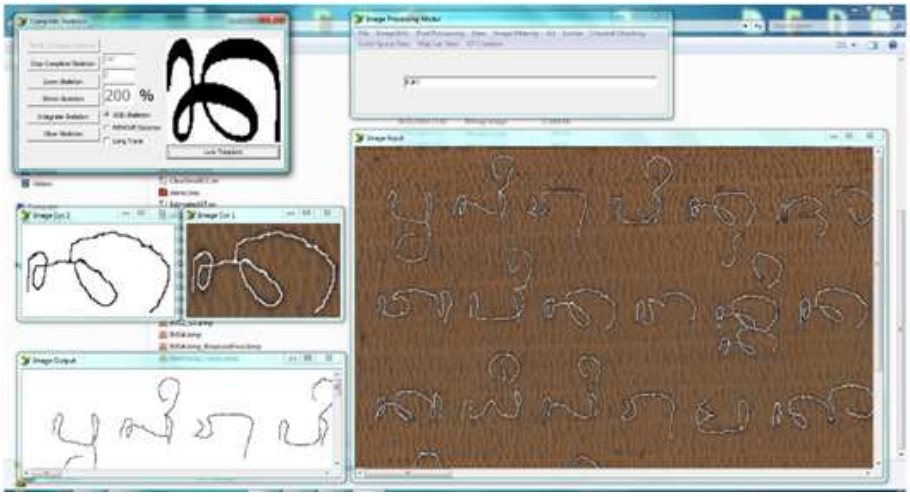


Figure 7.1: Snapshot of Prototype Interface used for manual correction of skeleton [1]

Figure 7.2: Estimated ground truth of a nondegraded palm leaf manuscript image [1]

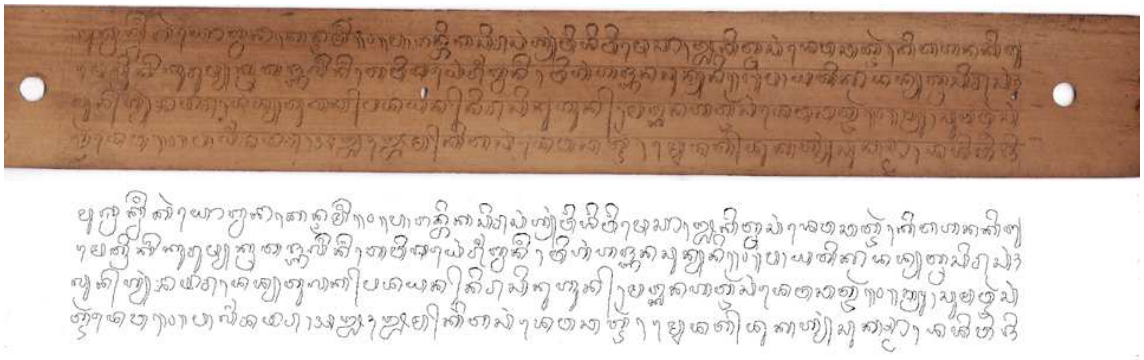


Figure 7.3: Original Image and the skeleton ground truth [1]

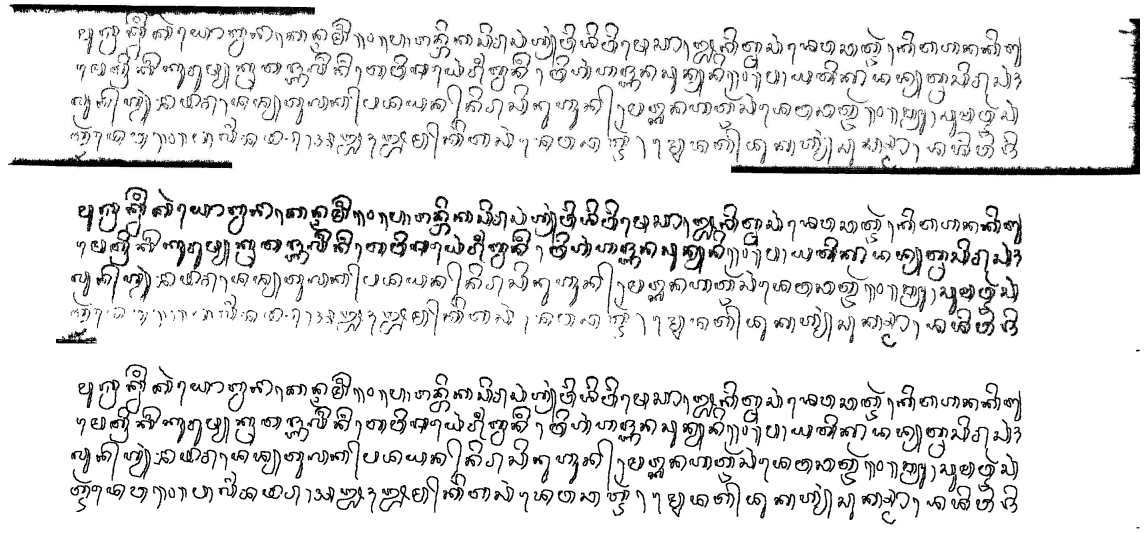


Figure 7.4: Ground truth image constructed with an initial binarized image of Niblack's method, Multi Resolution Otsu's method, and without any constraint of initial binarized image [1]

for RLS Method, and pruning the branch of 2 pixels. We performed the manual correction of the skeleton, guided by the transcription of the manuscript provided by a philologist to obtain the skeleton ground truth image of the manuscript. Figure 7.3 shows an example of a low quality palm leaf manuscript and the skeleton ground truth image. We first experimented with the construction of an estimated ground truth image by applying a constraint of Canny edge image and an initial binarized image. For example, we used the binarized image from Niblack's method or the multi resolution version of Otsu's method as the constraint. The estimated ground truth image really depends on the initial binarized image used as a constraint. We then experimented with the construction of the ground truth image without any initial binarized image as a constraint. The result is shown in Figure 7.4. Based on visual criteria, the proposed algorithm seems to achieve a better estimated ground truth image with respect to image quality and preservation of meaningful textual character information. Some other results of ground truth binarized images for degraded low quality palm leaf manuscripts are shown in Figure 7.5.

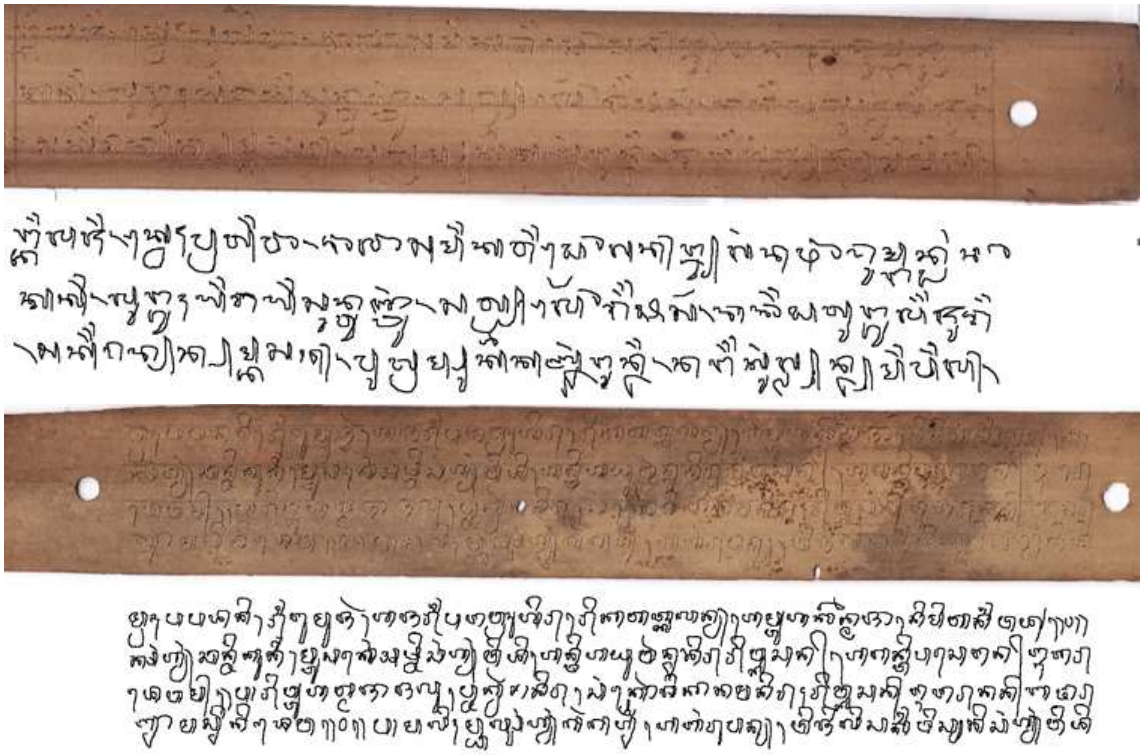


Figure 7.5: Two palm leaf manuscript images with their ground truth binarized images [1]

7.1.2.2 Experiments on the Effect of the Variation of Ground Truth Image

The ground truth images are normally accepted based on visual observation. It is unfortunately impossible to validate a ground truth construction methodology to create a perfect ground truth image from a real image. Recent studies on the analysis of binarization ground truthing [119] and the effect of ground truth on image binarization [120] discuss the influence human subjectivity during the manual correction process that may produce several different variations of ground truth. We design an experiment to compare the effect of the variation of ground truth image on the performance of the existing binarization methods. We generate two variations of our ground truth binarized image by applying an erosion and a dilation with a simple 3x3 square structuring element (Figure 7.6 and 7.7). Those two different versions of our ground truth image represent two possible human subjectivities on the border pixels of the character. We evaluate the performance of the binarization methods by using these three different versions of our ground truth binarized image. Table 7.1 shows the average value of F-Measure [12] of the binarization method for 9 palm leaf manuscripts. Our experiment shows that the rank performance of the binarization methods evaluated on three different versions of ground truth image is stable. This means that our original ground truth images are robust enough to be used to evaluate and to select future binarizations method for ancient palm leaf manuscripts. The low value of F-Measure for all binarization methods shows that the binarization problem is still an open question especially for specific document images such as our ancient palm leaf manuscripts.

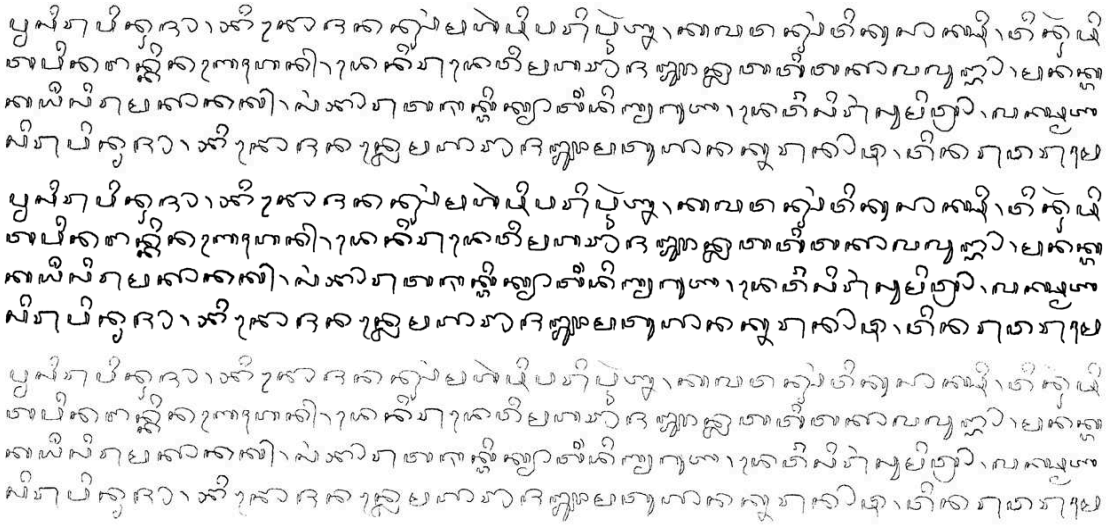


Figure 7.6: Two variations of ground truth binarized image: Original ground truth image, Dilated ground truth image and Eroded ground truth image

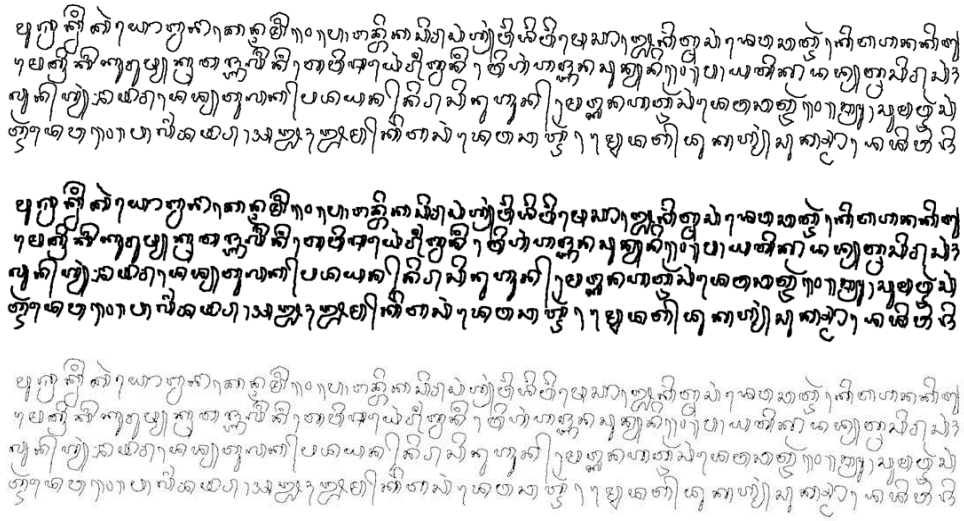


Figure 7.7: Two variations of ground truth binarized image: Original ground truth image, Dilated ground truth image and Eroded ground truth image

Table 7.1: Average F-Measure of existing binarization method evaluated on three versions of ground truth image created by our proposed scheme

Rank	Method	GT Original	GT Eroded	GT Dilated
1	NICK	60,48	44,78	63,47
2	SAUVOLA	57,61	40,62	61,39
3	WOLF	54,02	41,43	56,26
4	NIBLACK	48,35	29,47	54,63
5	HOWE	42,47	28,98	45,76
6	RAIS	34,61	18,82	41,68
7	OTSU	32,88	18,90	41,32

Table 7.2: Collection of palm leaf manuscripts from Museum Gedong Kirtya, Singaraja Bali, Indonesia

No	Content	Reference Code in Museum	Nb pages for experiment
1.	Sima Desa Tejakula	IIA-5-789	7
2.	Awig-awig Desa Tunju	IIA-10-1534	2
3.	Dewa Sasana	IIB-2-180	4
4.	Panugrahan Bhatara Ring Pura Pulaki	IIIB-12-306	8
5.	Buwana	IIIB-42-1526	2
6.	Pambadah	IIIB-45-2296	5
7.	Krakah Sang Graha	IIIC-19-1293	6
8.	Taru Pramana	IIIC-20-1397	7
9.	Siwa Kreket	IIIC-23-1506	6
		TOTAL	47

7.1.2.3 Analysis of Binarized Image Variability

Regarding the human intervention in the ground truthing process, the subjectivity effect on the construction of ground truth binarized images needs to be analyzed and reported. The work of [119] and [120] analyzed the binarization ground truthing and the effect of ground truth on image binarization of DIBCO binarized image dataset [114]. The study stated that the different choice of binarization ground truth affects the binarization algorithm design and the performance can vary significantly depending on the choice of ground truth.

In this sub section, we present an experiment in a real condition to analyze the human intervention subjectivity on the construction of ground truth binarized image and to measure quantitatively the ground truth variability of palm leaf manuscripts with different binarization evaluation metrics [18]. This experiment measures the difference between two ground truth binarized images from two different ground truthers. In this experiment, we adopted a semi-automatic framework for the construction of ground truth binarized images which is described in Sub Section 5.3.2.1. In order to measure the variability of human subjectivity in our ground truth creation, in this experiment, we did not apply any initial binarization and skeletonization methods. The skeletonization process is completely performed by human. The sample images used in this experiment are 47 images randomly selected from the palm leaf manuscript corpus of AMADI Project (Table 7.2).

7.1.2.3.1 Scheme of Experiment

For this experiment, 70 students were asked to trace manually the skeleton of the Balinese character found in palm leaf manuscripts with the PixLabeler tool [17] (Figure 7.8). One student worked with two different images, and one image was ground truthed by two different students. These two manually skeletonized images will be re-skeletonized with

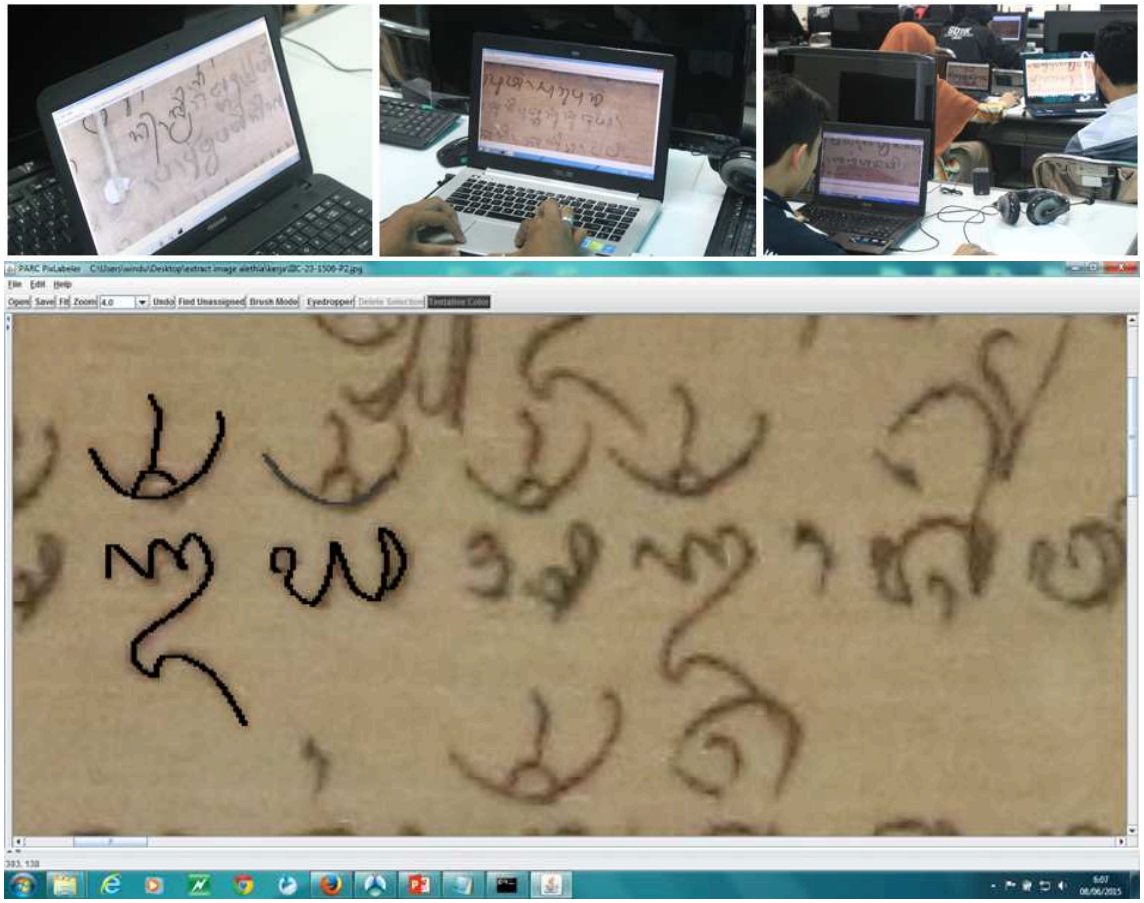


Figure 7.8: Students manually trace the skeleton of the Balinese character found in palm leaf manuscript with the PixLabeler tool [17]

Matlab's function bwmorph[1] to make sure that the skeleton is one pixel wide for the next process of automatic ground truth estimation with conditional dilation and Canny edge constraint. The final estimated ground truth binarized image is then automatically constructed by dilating the skeleton image, constrained by the character edges. The skeleton drawn manually by the user is dilated until the Canny edges intersect each binarized component of the dilated skeleton in a ratio of 0.1. This value of minimal ratio between the number of the pixels in the intersection of the Canny edge and the number of pixels of the dilated skeleton is found based on our empirical experiment and observation on the thickness of the character strokes in our manuscripts [18]. Figure 7.9 shows the scheme diagram of our experiment. Figure 7.10 shows some sample images as the result example of this experiment.

7.1.2.3.2 Measure of Ground Truth Variability

By observing visually the two skeletonized images created by two different ground truthers, we can see how different the results of the two ground truthers are in choosing the trace of the character skeleton. All the broken parts of an image of intersection between two skeletonized images show the different skeleton traces between two ground truthers. And all the double-lined parts in an image of union between two skeletonized images

[1] http://fr.mathworks.com/help/images/ref/bwmorph.html

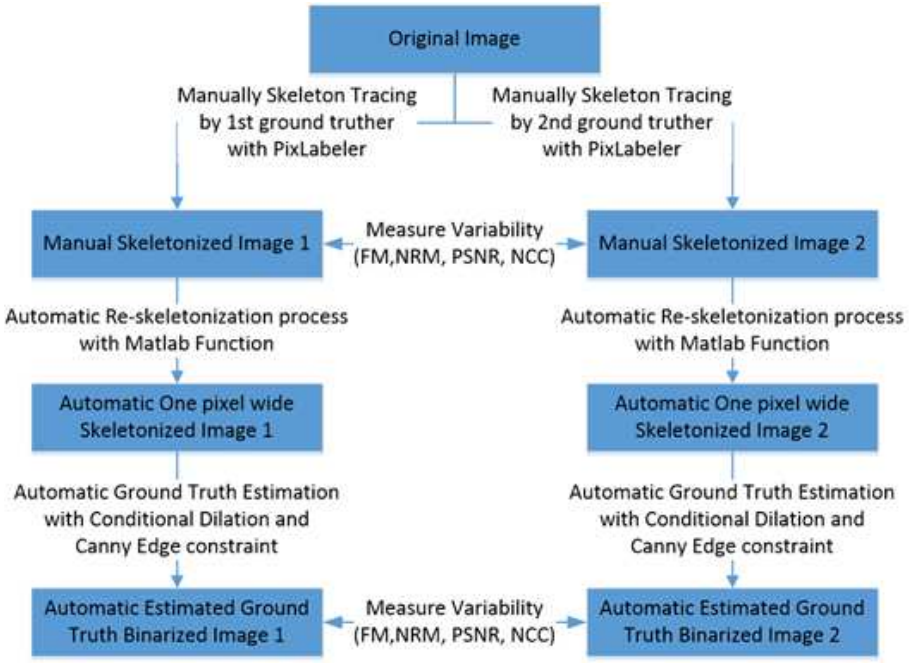


Figure 7.9: Scheme diagram of experiment [18]

Table 7.3: Variability between two manually skeletonized ground truthed image [18]

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	58,945	0,371	0,458	60,166
Minimum	14,058	0,209	0,209	26,882
Mean	41,459	0,302	0,303	33,196
Variance	77,764	0,002	0,003	60,083

show how different the positions of the skeleton traced by two ground truthers are.

First, we measure the variability between two skeletonized ground truthed images manually drawn by two different ground truthers (Table 7.3) [18]. The wide range between the maximum and the minimum value and also the mean and variance value of all three binarization evaluation metrics from 47 images show that there is a large variability between the ground truthers for each image. We then measured the variability between the two ground truth binarized images automatically estimated from two different manually skeletonized images for each image of a manuscript. Table 7.4 illustrates this variability [18]. The wide range between the maximum and the minimum value and also the mean and variance value of all three binarization evaluation metrics show that there is still a large variability between the estimated ground truth images for each image.

By comparing the value of binarization evaluation metrics between the two manually skeletonized ground truth images (Table 7.3) and between the two automatic estimated ground truth images (Table 7.4), we can see that the variability of two ground truth images in F Measure and NRM for all images decreases after the estimation ground truth process. The value of PSNR decreases because the number of different foreground-background pixels between the two estimated ground truth images also increases after

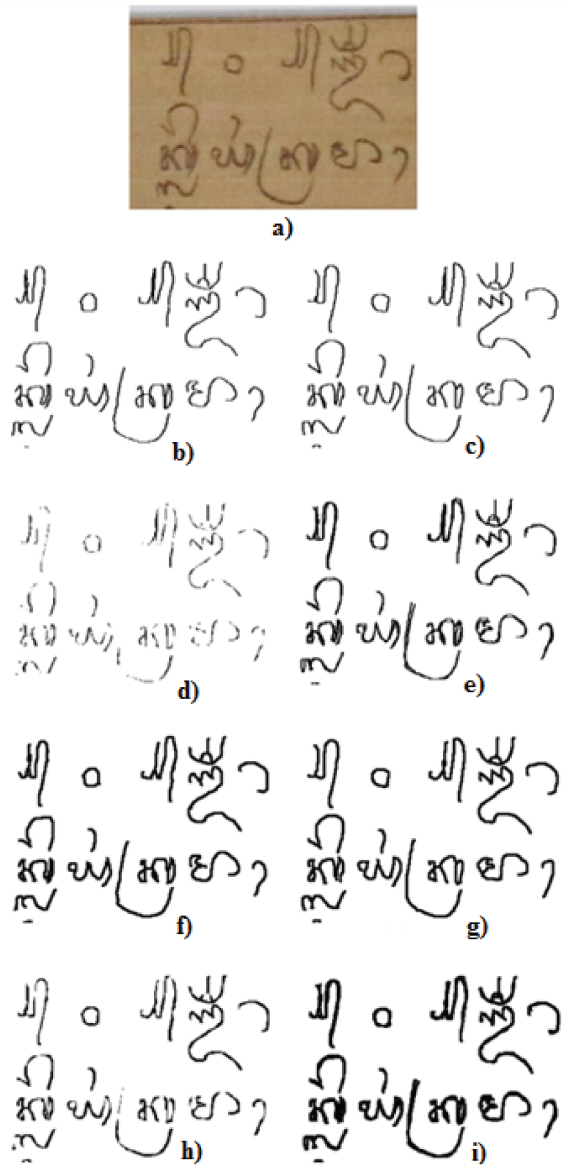


Figure 7.10: Example of ground truth binarized image from the experiment: (a) original image, (b) skeletonized image by 1st ground truther, (c) skeletonized image by 2nd ground truther, (d) image intersection between (b) and (c), (e) image union between (b) and (c), (f) estimated ground truth binarized image from (b), (g) estimated ground truth binarized image from (c), (h) image intersection between (f) and (g), (i) image union between (f) and (g) [18]

Table 7.4: Variability between two ground truthed image automatically estimated from two different manually skeletonized image [18]

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	74,731	0,309	0,446	59,196
Minimum	18,615	0,128	0,130	23,961
Mean	59,556	0,214	0,215	31,110
Variance	89,880	0,002	0,003	61,383

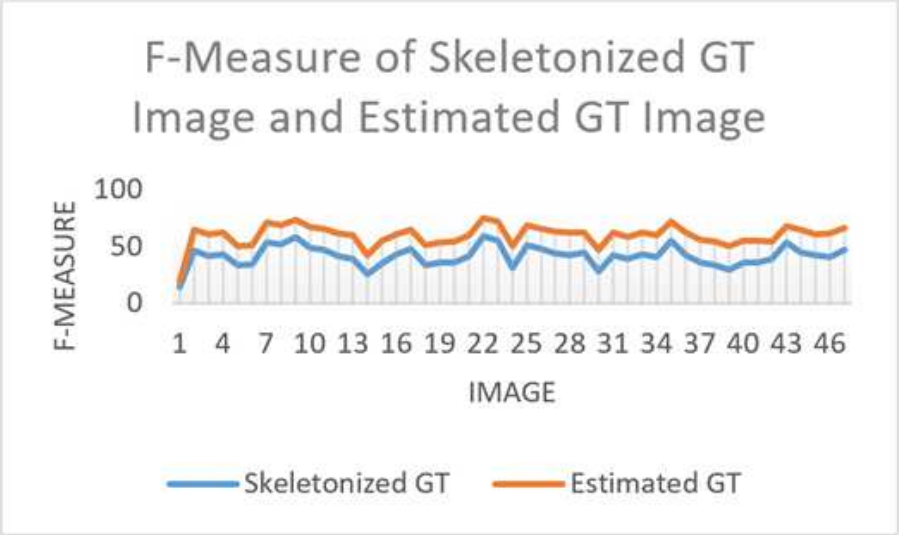


Figure 7.11: Comparison of F-Measure between two skeletonized ground truth image and between two estimated ground truth images [18]

Table 7.5: Variability between ground truth image estimated from union of two skeleton images with ground truth image estimated from the first ground truther [18]

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	89,758	0,076	0,418	67,095
Minimum	27,823	0,038	0,064	29,854
Mean	80,539	0,066	0,132	37,759
Variance	71,677	0,000	0,003	70,775

the automatic estimation process, not only the number of common foreground pixels from the two estimated ground truth images. Figures 7.11 to 7.14 show that the ground truth estimation process tends to decrease the variability between two ground truthers to produce a better match between two ground truth images.

We also tested and estimated the ground truth binarized image from the union of two skeleton images manually drawn by two differents ground truthers (see exemple in Figure 7.10(e)). The variability between this estimated union ground truth image with two other estimated ground truth images from each ground truther is then measured. Tables 7.5 and 7.6 illustrate the results of the comparison metric for all images in the experiments [18]. The ground truth image, estimated from the union of two skeleton images, indicates a better match with two other ground truth images from two different ground truthers.

Based on our data survey after the experiment with all ground truthers, we have observed and remarked some facts on the ground truth creation of palm leaf manuscripts as follow: The Balinese alphabets found on the manuscripts are not daily used by the ground truthers. Most of the ground truthers learned those alphabets in their elementary school until their junior or senior high school, but they never re-used those alphabets after the classroom learning process. There are some characters of the alphabet that they have never seen before. For those kinds of characters, the ground truthers could not make a smooth and natural trace of the character skeleton. Regarding the variability of

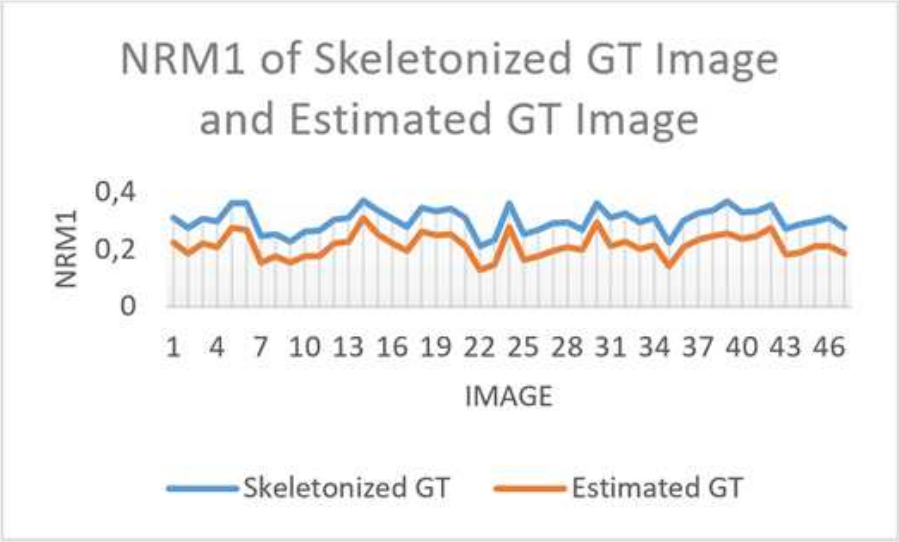


Figure 7.12: Comparison of NRM1 between two skeletonized ground truth image and between two estimated ground truth images [18]

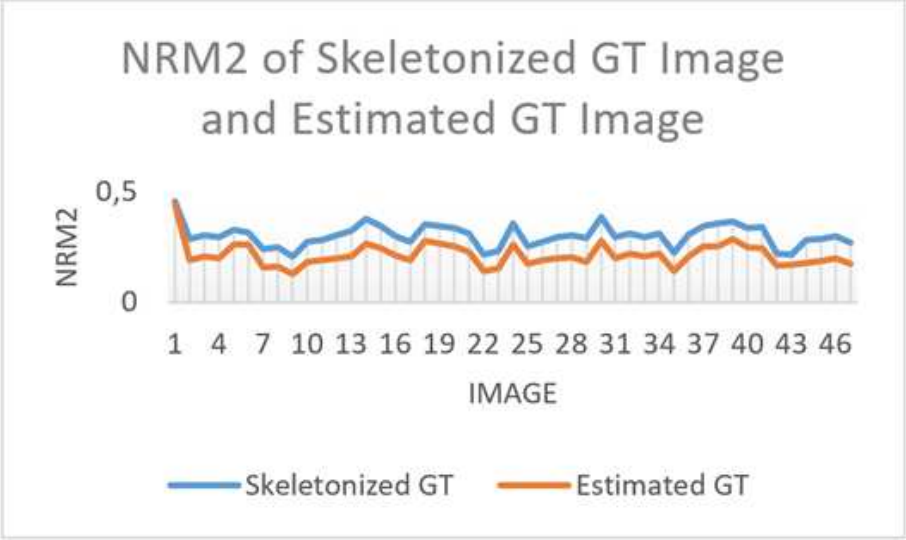


Figure 7.13: Comparison of NRM2 between two skeletonized ground truth image and between two estimated ground truth images [18]

Table 7.6: Variability between ground truth image estimated from union of two skeleton images with ground truth image estimated from the second ground truther [18]

Comparison metric	FM	NRM1	NRM2	PSNR
Maximum	94,182	0,090	0,227	65,188
Minimum	66,806	0,025	0,035	30,815
Mean	81,155	0,067	0,129	37,816
Variance	17,054	0,000	0,001	63,464

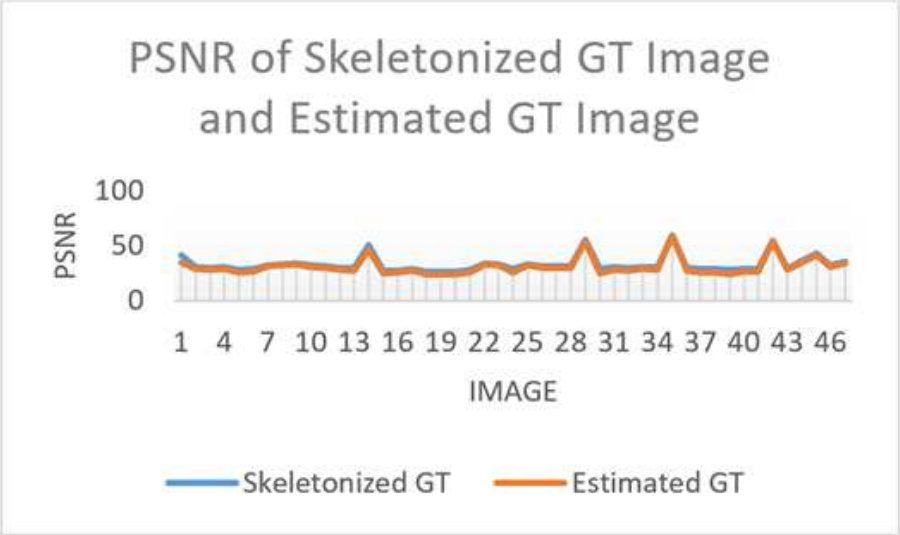


Figure 7.14: Comparison of PSNR between two skeletonized ground truth image and between two estimated ground truth images [18]

the ground truth images produced in this experiment, we suggest that this kind of important fact or condition should be always taken into account in every ground truthing process in a ancient manuscript project. The time needed to semi-manually correct the skeleton of an image from an initial automatic method can be much greater than making the skeleton totally manual by starting from zero. In our first trial experiment, we need between 4 and 6 hours to correct the semi-automatic generated skeleton. It is due to the physical carateristics of the manuscripts which make the binarizing and skeletonizing methods not produce the optimal good skeleton of the characters. We finally decided to make it totally manual, and it takes between 2 and 3 hours to trace the skeleton starting from zero.

From the result of this experiment, we proved that human subjectivity has a great effect in producing a great variability of ground truth binarized image. This phenomenon becomes much more visible when we are working on the binarization process of ancient documents or manuscripts where the physical characteristics and conditions of the manuscript are not good enough or it is still hard to be ground truthed even by a human.

7.1.3 Binarization Methods Evaluation: Results and Discussion

After generating the binarized ground truth images, we tested and evaluated some well-known standard binarization methods. We also tested the binarization methods from the DIBCO competition [114, 2] for example the Howe's method [10] and the methods from the binarization challenge of the ICFHR competition[2] [23].

The experimental results for the binarization task are presented in Table 7.7. These results show that the performance of all methods on each dataset are still quite low. Most of the methods only achieve less than 50% FM score. It means that palm leaf manuscripts are still an open challenge for the binarization task. The different parameter values for the local adaptive binarization methods gives actually a significant improvement in the

[2]amadi.univ-lr.fr/ICFHR2016_Contest

performance of each other, but it is still an unsatisfactorily result. In these experiments, the ICFHR G1 method was evaluated for the Khmer and Sundanese datasets by using the pre trained Balinese training set weighted model. Based on these experiments, Niblack's method gives the highest FM score for Sundanese manuscripts (Figure 7.15), the ICFHR G1 method gives the highest FM score for the Khmer manuscripts (Figure 7.16), and the ICFHR G2 gives the highest FM score for the Balinese manuscripts (Figure 7.17). Visually there are still many broken and unrecognizable characters/glyphs and noise is detected in the images.

Table 7.7: Experimental results for binarization task in F-Measure (FM), Peak SNR (PSNR), and Negative Rate Metric (NRM). A higher F-measure and PSNR, and a lower NRM, indicate a better result

Methods	Parameter	Manuscripts	FM (%)	NRM	PSNR (%)
OtsuGray [5, 6]	Otsu from gray image Using Matlab graythresh[3]	Balinese	18.98178	0.398894	5.019868
		Khmer	23.92159	0.313062	7.387765
		Sundanese	23.70566	0.326681	9.998433
OtsuRed [5, 6]	Otsu from red image channel Using Matlab graythresh	Balinese	29.20352	0.300145	10.94973
		Khmer	21.15379	0.337171	5.907433
		Sundanese	21.25153	0.38641	12.60233
Sauvola [5, 7] [6, 8, 4, 9]	window = 50, k = 0.5, R = 128	Balinese	13.20997	0.462312	27.69732
		Khmer	44.73579	0.268527	26.06089
		Sundanese	6.190919	0.479984	**24.78595**
Sauvola2 [5, 7] [6, 8, 4, 9]	window = 50, k = 0.2, R = 128	Balinese	40.18596	0.274551	25.0988
		Khmer	47.55924	**0.155722**	21.96846
		Sundanese	43.04994	0.299694	23.65228
Sauvola3 [5, 7] [6, 8, 4, 9]	window = 50, k = 0.0, R = 128	Balinese	35.38635	0.165839	17.05408
		Khmer	30.5562	0.190081	12.78953
		Sundanese	40.29642	0.181465	16.25056
Niblack [5, 7] [6, 8, 4]	window = 50, k = -0.2	Balinese	41.55696	0.175795	21.24452
		Khmer	38.01222	0.160807	16.84153
		Sundanese	**46.79678**	0.195015	20.31759
Niblack2 [5, 7] [6, 8, 4]	window = 50, k = 0.0	Balinese	35.38635	0.165839	17.05408
		Khmer	30.5562	0.190081	12.78953
		Sundanese	40.29642	0.181465	16.25056
Nick [4]	window = 50, k= -0.2	Balinese	37.85919	0.328327	27.59038
		Khmer	51.2578	0.176003	24.51998
		Sundanese	29.5918	0.390431	24.26187
Rais [5]	window = 50	Balinese	34.46977	0.171096	16.84049
		Khmer	31.59138	0.187948	13.52816
		Sundanese	40.65458	**0.177016**	16.35472
Wolf [8, 4]	window = 50, k = 0.5	Balinese	27.94817	0.392937	27.1625
		Khmer	46.78589	0.23739	25.1946
		Sundanese	42.40799	0.299157	23.61075
Howe1 [10]	Default values	Balinese	44.70123	0.267627	28.35427
		Khmer	40.20485	0.280604	25.59887
		Sundanese	45.90779	0.235175	21.90439
Howe2 [10]	Default values	Balinese	40.5555	0.273994	28.02874

continued on next page

Figure 7.15: Binarization of Sundanese manuscript with Niblack's method

Figure 7.16: Binarization of Khmer manuscript with ICFHR G1 method

Table 7.7 – continued from previous page

Methods	Parameter	Manuscripts	FM (%)	NRM	PSNR (%)
		Khmer	32.35603	0.294016	25.96965
		Sundanese	35.35973	0.274865	22.36583
Howe3 [10]	Default values	Balinese	42.15377	0.304962	28.38466
		Khmer	30.7186	0.382087	26.36983
		Sundanese	25.77321	0.350349	23.66912
Howe4 [10]	Default values	Balinese	45.73681	0.273018	28.60561
		Khmer	36.48396	0.280519	25.83969
		Sundanese	38.98445	0.281118	22.83914
ICFHR G1	See ref. [23]	Balinese	63.32	0.15	31.37
		Khmer	**52.65608**	0.250503	**28.16886**
		Sundanese	38.95626	0.329042	24.15279
ICFHR G2	See ref. [23]	Balinese	**68.76**	**0.13**	**33.39**
		Khmer	-	-	-
		Sundanese	-	-	-
ICFHR G3	See ref. [23]	Balinese	52.20	0.18	26.92
		Khmer	-	-	-
		Sundanese	-	-	-
ICFHR G4	See ref. [23]	Balinese	58.57	0.17	29.98
		Khmer	-	-	-
		Sundanese	-	-	-

7.1.4 Conclusions

For the semi automatic ground truth construction, the specific binarization scheme proposed in this work provides a good initial image of skeleton with respect to image quality

Figure 7.17: Binarization of Balinese manuscript with ICFHR G2 method

and preservation of meaningful textual character information. This method achieves a better ground truth for degraded low quality palm leaf manuscripts. This scheme adapts and performs better in constructing the ground truth of binarized images for palm leaf manuscripts. The ground truth images constructed with this scheme are robust enough to be used to evaluate and to select future optimal binarization methods for ancient palm leaf manuscripts.

Human subjectivity has a great effect in producing a great variability of ground truth binarized image. This phenomen becomes much more visible when we are working on the binarization process of ancient documents or manuscripts where the physical characteristics and conditions of the manuscript are not good enough or it is still hard to be ground truthed even by human. The choice of ground truth data set and the variability within the ground truth should be analyzed quantitatively before the performance measure of any binarization methods. In the case of a manuscript with specific ancient characters, the qualitative observation and validation should also be made by the philologist to guarantee the correctness of the binarized characters on the manuscripts.

Binarizing the palm leaf manuscript images seems very challenging. The low value of F-Measure for all binarization methods shows that the binarization problem is still an open question especially for specific document images such as our ancient palm leaf manuscripts. Still, with many broken and unrecognizable characters/glyphs and noises detected in the images, binarization should be reconsidered the first step in the DIA process for palm leaf manuscripts. On the other hand, although there are already training-based DIA methods that do not require this binarization process, they usually require adequate training data.

7.2 Text Line Segmentation Evaluation

7.2.1 Evaluation Metrics

We use the evaluation criteria and tool provided by the ICDAR2013 Handwriting Segmentation Contest[4] [121] (Figure 7.18). First, the one-to-one (***o2o***) match score is computed for a region pair based on the evaluator's acceptance threshold. To calculate the

[4]http://users.iit.demokritos.gr/~nstam/ICDAR2013HandSegmCont/Protocol.html

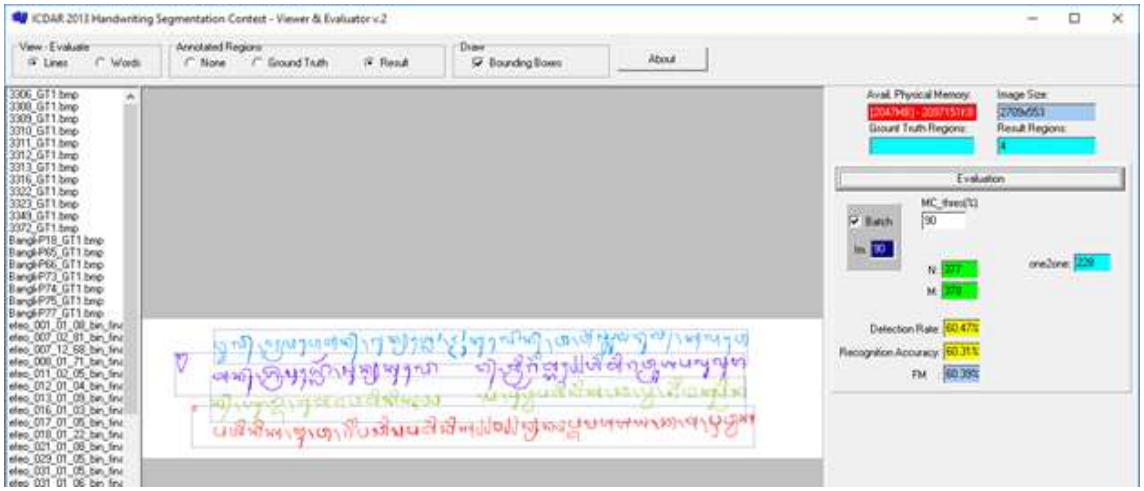


Figure 7.18: ICDAR2013 Handwriting Segmentation Contest – Viewer and Evaluator

(***o2o***) match score, the MatchScore table is first calculated according to the intersection of the pixel sets of the result and the ground truth[5]. Let G_i be the set of all points of the i^{th} ground truth regions, R_j the set of all points of the j^{th} result region, *T(s)* a function that counts the elements of set *s*. *MatchScore(i,j)* represents the matching results of i^{th} ground truth region and the j^{th} result region as

$$MatchScore(i,j) = \frac{T(G_i \cap R_j)}{T(G_i \cup R_j)}. \tag{7.9}$$

A pair of ground truth and result regions is a one-to-one match (o2o) if the matching score for this pair is equal to or above the evaluator's acceptance threshold [72]. In our experiments, we used 90% as the acceptance threshold. A g_one-to-many match (go2m) is a ground truth text line that "partially" matches with two or more text lines in the detected result. A g_many-to-one match (gm2o) corresponds to two or more ground truth text lines that "partially" match with one detected text line (or word). A d_one-to-many match (do2m) is a detected text line that "partially" matches two or more text lines in the ground truth. Finally, a d_many-to-one match (dm2o) corresponds to two or more detected text lines that "partially" match one text line in the ground truth.

Let ***N*** be the count of ground truth elements, ***M*** be the count of result elements, and w_1, w_2, w_3, w_4, w_5, w_6 are pre-determined weights. With the ***o2o***, ***go2m***, ***gm2o***, ***do2m***, and ***dm2o*** score, three metrics, detection rate (***DR***), recognition accuracy (***RA***), and performance metric (***FM***), are calculated as described in [72, 122]:

$$DR = w_1 \frac{o2o}{N} + w_2 \frac{go2m}{N} + w_3 \frac{gm2o}{N}. \tag{7.10}$$

$$RA = w_4 \frac{o2o}{N} + w_5 \frac{do2m}{N} + w_6 \frac{dm2o}{N}. \tag{7.11}$$

$$FM = \frac{2 * DR * RA}{DR + RA}. \tag{7.12}$$

[5] http://users.iit.demokritos.gr/~nstam/ICDAR2013HandSegmCont/Evaluation.html

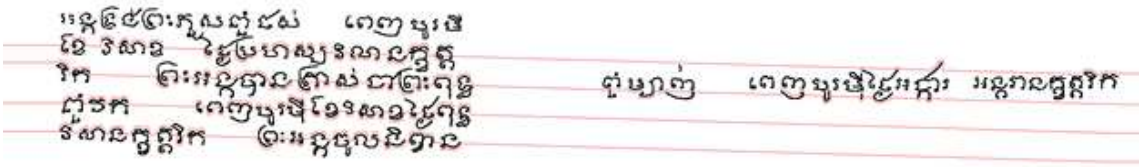


Figure 7.19: The A* Path Planning approach is failed to determine the starting state of each line in some manuscripts

7.2.2 Results and Discussion

For the evaluation of the text line segmentation methods, we performed the experiments in two parts. In the first part, we performed the evaluation test for five text line segmentation methods only for smaller dataset (Bali 1, Khmer 1 and Sunda 1). In the second part, we performed the experiments with more complete and bigger dataset.

7.2.2.1 Experiment Part 1

In the first part, we performed the evaluation test for five text line segmentation methods (APP, Shredding, Seam Carving, ALCM and A* Path Planning) for dataset Bali 1, Khmer 1 and Sunda 1. Table 7.8 shows the results of performance evaluation of each method on each collection of manuscripts. We also calculated the performance of each method on the total collection. In general, the methods which are applied to binary images achieved a good enough result. It is because we performed our experiments on the binary ground truth image of the manuscripts. These images are the ideal condition of the binary images that we can expect from the original manuscripts. The performance of binary image based methods will be greatly influenced by the quality of the binary image.

In our experiments, without doing any post processing task, the A* Path Planning approach achieves the best results for all manuscript collections. Nevertheless, this approach greatly depends on the result of binarization and projection profile analysis in localizing the text line and the starting state of each line. This approach did not perform well when it is failed to detect the starting points of lines, because the manuscript contains some short lines (Figure 7.19).

The APP method performs better on the collection of Khmer manuscripts because Khmer manuscripts often contain some spaces between words or between some shorter sub text lines. This characteristic fits well with the APP method as it divides the manuscripts into some smaller vertical zones. The APP approach greatly relies on information extracted from the global horizontal projection profile at the beginning of the process to extract some important referenced values such as the number of text lines, the average line position of each text line, and also the average height of text lines. Most of the rules which are applied in the next step of the APP approach depend on the spatial information of the text line provided by those referenced values. For example the average height of the text lines is used to detect the base line of the vowel, and the number of text lines is used to check (to insert or delete) the correct base line to each column. If those values are incorrect, the accuracy of the approach will drop significantly. It therefore does not work well with document images with skewed and curved or fluctuating text lines. The APP method did not perform well on the collection of Balinese and Sundanese manuscripts because the size of the ascenders and descenders of the characters in Balinese and Sundanese script is almost the same with the size of the character itself and it occupies the

Table 7.8: Experimental results for the text line segmentation task: the count of ground truth elements (N), and the count of result elements (M), the one-to-one (o2o) match score is computed for a region pair based on 90% acceptance threshold, detection rate (DR), recognition accuracy (RA), and performance metric (FM)

Method	Collection	N	M	o2o	DR(%)	RA(%)	FM(%)
APP (binary)	Bali 1	140	168	100	71.42	59.52	64.93
	Sunda 1	46	51	32	69.56	62.74	65.97
	Khmer 1	191	207	164	85.86	79.22	82.41
	All Collection	377	426	296	78.51	69.48	73.72
Shredding (binary)	Bali 1	140	167	123	87.85	73.65	80.13
	Sunda 1	46	142	25	54.34	17.60	26.59
	Khmer 1	191	185	91	47.64	49.18	48.40
	All Collection	377	494	239	63.39	48.38	54.87
Energy Function Based Shredding (binary)	Bali 1	140	178	128	91.42	71.91	80.50
	Sunda 1	46	50	46	100.0	92.00	95.83
	Khmer 1	191	190	181	94.76	95.26	95.01
	All Collection	377	418	355	94.16	84.92	89.30
Basic Seam Carving (binary)	Bali 1	140	820	73	52.14	8.90	15.20
	Sunda 1	46	137	13	28.26	9.48	14.20
	Khmer 1	191	205	0	0	0	0
	All Collection	377	1162	86	22.81	7.40	11.17
Basic Seam Carving (grayscale)	Bali 1	140	1087	80	57.14	7.35	13.03
	Sunda 1	46	172	14	30.43	8.13	12.84
	Khmer 1	191	214	1	0.52	0.46	0.49
	All Collection	377	1473	95	25.19	6.44	10.27
Complete Seam Carving (binary)	Bali 1	140	143	131	93.57	91.60	92.57
	Sunda 1	46	46	46	100.0	100.0	100.0
	Khmer 1	191	189	51	26.70	26.98	26.84
	All Collection	377	378	228	60.47	60.31	60.39
Complete Seam Carving (grayscale)	Bali 1	140	167	128	91.42	76.64	**83.38**
	Sunda 1	46	43	36	78.26	83.72	**80.89**
	Khmer 1	191	145	57	29.84	39.31	**33.92**
	All Collection	377	355	221	58.62	62.25	**60.38**
ALCM (grayscale)	Bali 1	140	322	20	14.28	6.21	8.65
	Sunda 1	46	66	4	8.69	6.06	7.14
	Khmer 1	191	392	59	30.89	15.05	20.24
	All Collection	377	780	83	22.01	10.64	14.34
A* Path Planning (binary)	Bali 1	140	141	137	97.85	97.16	**97.50**
	Sunda 1	46	46	46	100.0	100.0	**100.0**
	Khmer 1	191	190	182	95.28	95.78	**95.53**
	All Collection	377	377	365	96.81	96.81	**96.81**

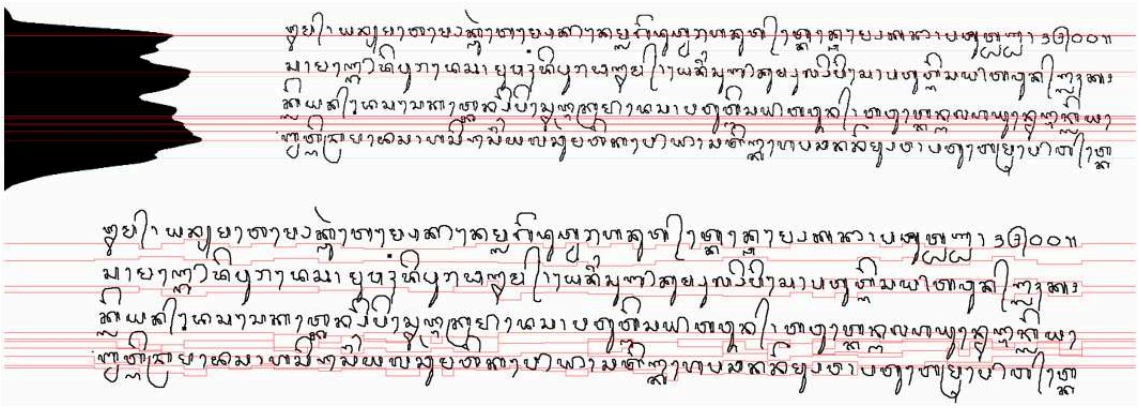


Figure 7.20: The rules in APP approach are failed to determine the base line of a Balinese manuscript

space between two consecutive text lines. The APP approach considers it as a new base line (Figure 7.20). The APP technique also has some problems caused by the components that are too far away from their main lines.

The shredding method achieved the best performance on Balinese manuscripts. The Balinese manuscripts normally have a wider space between two consecutive text lines than the Sundanese and Khmer manuscripts, so the shredding function works easier to separate the inter-text line areas. The use of our ellipse energy function significantly improves the performance of the recursive tracing function from the shredding method. For the Khmer manuscripts our proposed energy function is very optimal to force the energy transfer in one text line while preventing the energy transfer between two text lines (Figure 7.21).

For the seam carving based method, without any preprocessing and post processing task, the performance of this method is not very optimal, especially on the collection of Khmer manuscripts. The spaces between some shorter sub text lines influences greatly the minimum separating seam path. By passing these spaces, the separating seam paths jumped to other intra-text line areas or joined together into a single separating seam path (Figure 7.22). This behavior makes the seam carving method totally fail to separate the text lines. The complete scheme of seam carving gives a significant improvement in separating seam detection (Figure 7.23), but it greatly depends on the previous step of medial seam detection as a constraint which is based on the projection profile matching approach. If this first step fails to detect the correct medial axis of the textline, the separating seams will not be detected correctly (Figure 7.24).

7.2.2.2 Experiment Part 2

In the second part, we performed the evaluation test for the Seam Carving and the Adaptive Path Finding method for all datasets. The experimental results for the text line segmentation task are presented in Table 7.9. According to these results, both methods perform sufficiently well for most datasets except Khmer 1 (Figure 7.25-7.27). This is because all images in this set are of low quality due to the fact that they are digitized from microfilms. Nevertheless, the adaptive path finding method proves to achieve better results than the seam carving method on all datasets of palm leaf manuscripts in our experiment. The main difference between these two approaches is that instead of finding an optimal separating path within an area constrained by medial seam locations of two ad-

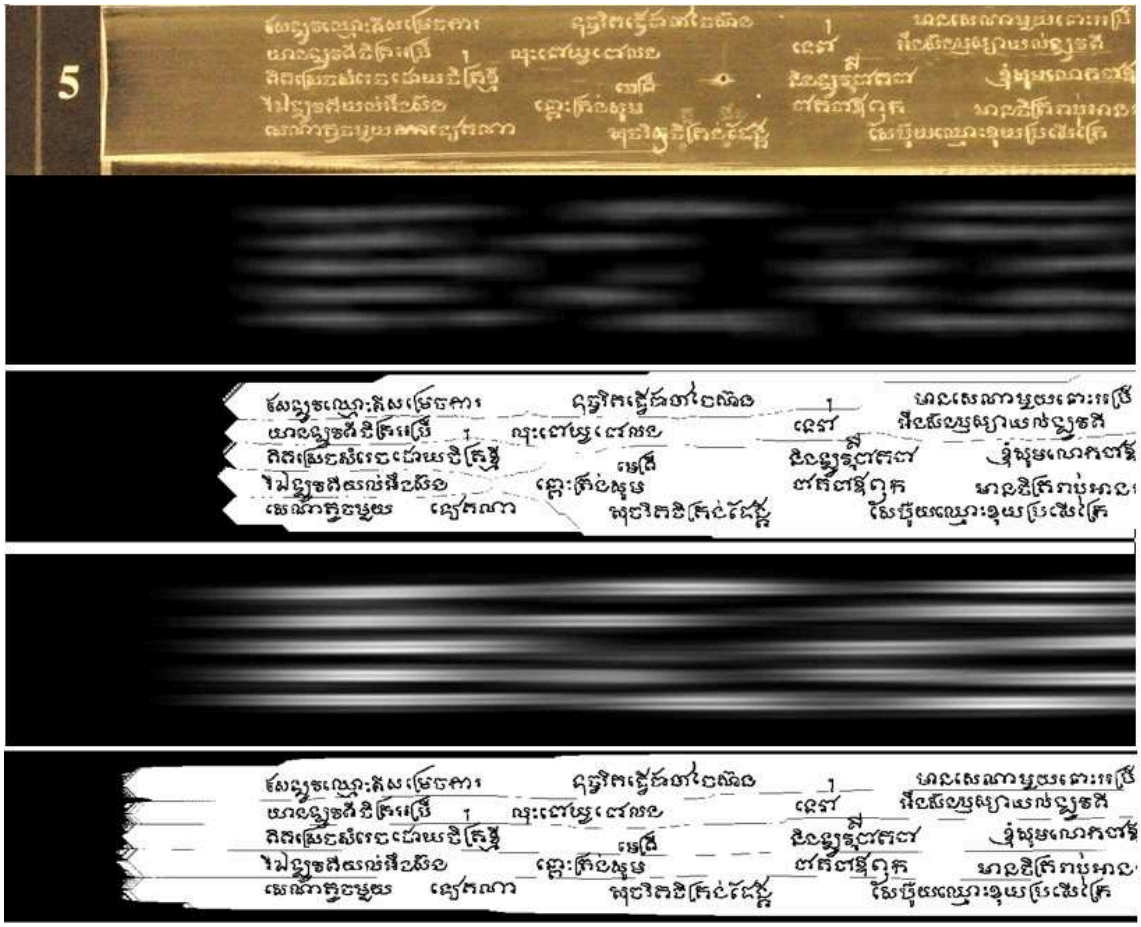


Figure 7.21: The improved energy function for shredding method. From top to bottom: original image of manuscript, original blurring function, detected line areas, improved energy function, improved detected line areas

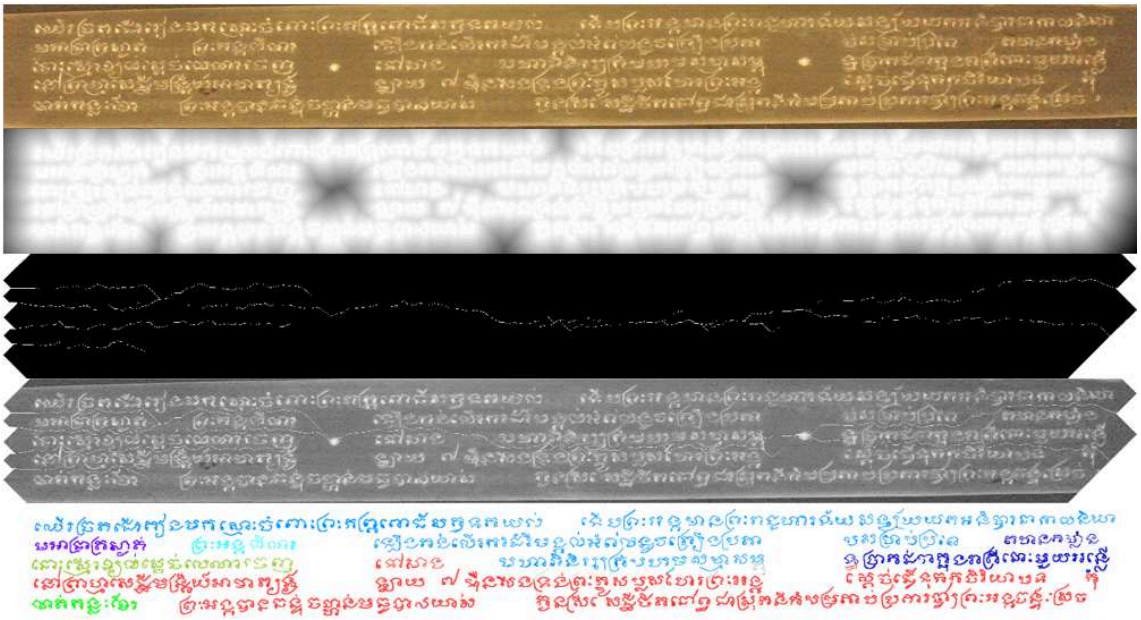


Figure 7.22: An example of the jumped and joined separating seam paths on a Khmer manuscript

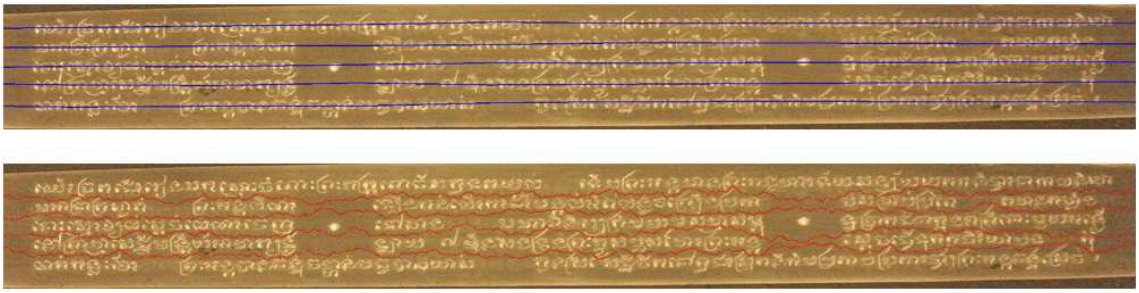


Figure 7.23: The medial seams (top) and the separating seams (bottom) of the manuscript of Figure 7.22 are correctly detected

Figure 7.24: The medial seams (top) and the separating seams (bottom) of the manuscript are not correctly detected

jacent lines (in the seam carving method), the adaptive path finding approach tries to find a path close to an estimated straight seam line section. These line sections already represent well enough the seam borders between two neighboring lines, so they can be considered as a better guide for finding good paths hence producing better results.

One common error that we encounter for both methods is in the medial position computation stage. Detecting correct medial positions of text lines is crucial for the path finding stage of the methods. In our experiment, we noticed that some parameters play an important role. For instance, the number of columns/slices r of the seam carving method and the high and low thresholding values of the edge detection algorithm in the adaptive path finding approach. In order to select these parameters, a validation set consisting of five random pages is used. The optimal values of the parameters are then empirically selected based on the results from this validation set.

7.2.3 Conclusions

We investigated the performance of six text line segmentation methods by conducting the comparative experimental studies for the collection of palm leaf manuscript images. Three methods work on the binary image: the Adaptive Partial Projection (APP) line segmentation approach, the A* Path Planning approach, and the shredding method (with

Figure 7.25: Text line segmentation of Balinese manuscript with the Seam Carving method (green) and Adaptive Path Finding (red)

Table 7.9: Experimental results for the text line segmentation task: the count of ground truth elements (N), and the count of result elements (M), the one-to-one (o2o) match score is computed for a region pair based on 90% acceptance threshold, detection rate (DR), recognition accuracy (RA), and performance metric (FM)

Methods	Manuscripts	N	M	o2o	DR (%)	RA (%)	FM (%)
Seam carving [82]	Balinese 1	140	167	128	91.42	76.64	83.38
	Bali-2.1	181	210	163	90.05	77.61	83.37
	Bali-2.2	182	219	161	88.46	73.51	80.29
	Khmer 1	191	145	57	29.84	39.31	33.92
	Khmer 2	476	665	356	53.53	74.79	62.40
	Khmer 3	971	1046	845	87.02	80.78	83.78
	Sundanese 1	46	43	36	78.26	83.72	80.89
	Sundanese 2	242	257	218	90.08	84.82	87.37
Adaptive Path Finding [64]	Balinese 1	140	143	132	94.28	92.30	**93.28**
	Bali-2.1	181	188	159	87.84	84.57	**86.17**
	Bali-2.2	182	191	164	90.10	85.86	**87.93**
	Khmer 1	191	169	118	61.78	69.82	**65.55**
	Khmer 2	476	484	446	92.15	93.70	**92.92**
	Khmer 3	971	990	910	93.71	91.91	**92.80**
	Sundanese 1	46	50	41	89.13	82.00	**85.41**
	Sundanese 2	242	253	222	91.73	87.74	**89.69**

Figure 7.26: Text line segmentation of Khmer manuscript with the Seam Carving method (green) and Adaptive Path Finding (red)

Figure 7.27: Text line segmentation of Sundanese manuscript with the Seam Carving method (green) and Adaptive Path Finding (red)

our proposed energy function for shredding method). Three other methods can be directly applied to grayscale images: the Adaptive Local Connectivity Map (ALCM) method, the seam carving based method and Adaptive Path Finding. The results show that each method performed well on some specific characteristic of manuscript collection. The behavior of some methods is greatly influenced by some challenges which are clearly present on each collection of the Southeast Asian manuscripts. For example the A* Path Planning method did not perform well when the manuscript contains some short lines, because it is failed to detect the starting points of some lines. The APP method performs better on the manuscripts which contain some spaces between words or between some shorter sub text lines as it divided the manuscripts into some smaller vertical zones, but it did not work well with document image with skewed and curved or fluctuated text lines. The APP method did not perform well on the manuscripts with many ascenders and descenders of the character which occupy the space between two consecutive text lines. It suffers also from the components that are too far away from their main lines. The shredding method achieved the best performance on manuscripts which have normally a wider space between two consecutive text lines because the shredding function works easier to separate the inter-text line areas. For the seam carving based method, the spaces between some shorter sub text lines influence greatly the minimum separating seam path as it jumped to other intra-text lines area or joined together into a single separating seam path. This behavior can make the seam carving method totally fail to separate the text lines.

7.3 Isolated Glyph Recognition Evaluation

7.3.1 Evaluation Metrics

Following the evaluation method from the ICFHR competition [23], this process is evaluated with the recognition rate, i.e., the percentage of correctly classified samples over the test samples, C/N. Here C is the number of correctly recognized samples, and N is the total number of test samples.

7.3.2 Results and Discussion

For the evaluation of the isolated glyph recognition methods, we performed the experiments in three parts. In the first part, we performed the evaluation for 10 feature extraction methods. For the second part, the training based method with a neural network and unsupervised feature learning was investigated for the global and categorized glyph recognizer. In the third part, we conduct a broader evaluation of the robustness of the methods previously tested on Balinese script, to the other two different palm leaf manuscripts with Khmer and Sundanese scripts.

7.3.2.1 Experiment Part 1

In the first part, the experimental study on feature extraction methods for character recognition of Balinese script was performed [14]. We present this experimental study on feature extraction methods for character recognition of Balinese script on palm leaf manuscripts [14]. We investigated and evaluated the performance of 10 feature extraction methods and the proposed combination of features in 29 different schemes. For all experiments, a set of image patches containing Balinese characters from the original manuscripts are used as input, and a correct class of each character should be identified as a

result.

We used k=5 for the k-NN classifier, and all images are resized to 50x50 pixels (the approximate average size of characters in the collection), except for Gradient features where images are resized to 81x81 pixels to get evenly 81 blocks of 9x9 pixels, as described in [89]. For the Projection Histogram, features from horizontal projection histograms are combined with the features from vertical projection histograms to produce the final feature vectors of size 100. For the Celled Projection, features from horizontal celled projection are combined with the features from vertical celled projection with the cell width of 10 pixels to produce the final feature vectors of size 500. For the Distance Profile, the distance profiles are calculated for each row and column from four sides: left, right, top and bottom of the image to produce the final feature vectors of size 200. For the Crossing, features from horizontal crossing are combined with the features from vertical crossing for each row and column to produce the final feature vectors of size 100. For the Zoning, we computed zoning for binary image and grayscale image with 7 zone types (zone width or zone size = 5 pixels) and combined them into a 205 dimension feature vector. We also tested the zoning feature on the image of the skeleton. For the Moments, 7 Hu's moments are computed in 2 vertical zones and 2 horizontal zones with width 25 pixels, and 4 square zones of size 25x25 pixels to produce the final feature vectors of size 56. For the Kirsch Directional Edges, we computed Kirsch feature from the grayscale image with 25 smaller regions to produce a 100 dimension feature vector. Based on the empirical tests for our dataset, the Kirsch edge image can be properly thresholded with a threshold value of 128. The feature value is then normalized by the maximum value of edge pixel frequency from all regions. For the NPW, we computed NPW features in level 3 neighborhood with 25 smaller regions (N=25) to produce a 100-dimension feature vector. The feature value is normalized by the maximum value of average weight from all regions. We tested the performance of NPW features for both binary and grayscale images. For the HoG, we used the HoG implementation of VLFeat[6]. We computed HoG features from grayscale images with cell size of 6 pixels and with 9 different orientations to produce a 1984 dimension feature vector.

For the CNN, we used the Tensorflow library[7]. The first architecture of our network is illustrated in Figure 7.28. Given an input grayscale image of 28x28 pixels, the first convolutional layer (C1) is computed by using a sliding window of 5x5 pixels. We obtain C1 layer containing 28x28x32 neurons, where 32 is the number of feature maps chosen. For each sliding window on the neuron (i, j), the convolutional C(i, j) output can be computed by:

$$C(i,j) = \sum_{k=0}^{4} \sum_{l=0}^{4} (W_{(k,l)} I_{(i+k,j+l)}) \tag{7.13}$$

where W is a 5x5 matrix to be used as the shared weights matrix, and I is the input neuron. A rectified linear unit is then applied. We obtain: C=ReLu(b+C), where b is the bias. Then, we apply max-pooling using a window of 2x2 and choose the maximum activation in the selected region as the output. We obtain the P2 layer which consists of 14x14x32 neurons. After computing the second convolutional layer (C3) and second max-pooling (P4), we obtain a layer (P4) of 7x7x64 neurons. We add a fully-connected

[6]http://www.vlfeat.org/api/hog.html
[7]http://arxiv.org/abs/1603.04467

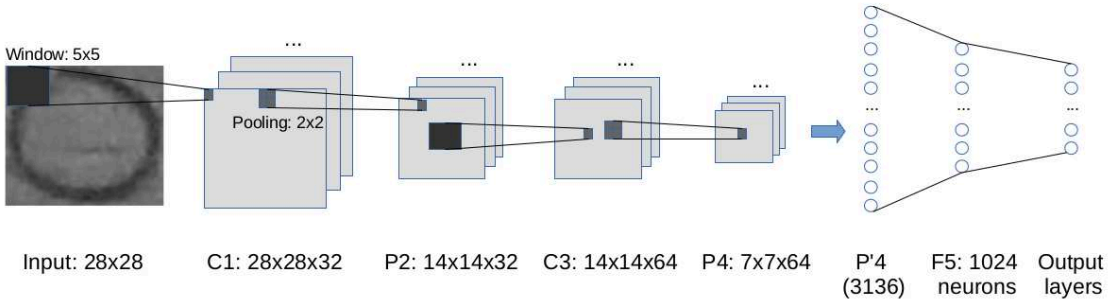


Figure 7.28: Architecture of multilayer convolutional neural network

layer (F5) of 1024 neurons, where

$$F5 = ReLU(P_4^{'}W_4 + b_4) \tag{7.14}$$

$P^{'}{}_4$ is a one dimension matrix containing the 3136 neurons in P_4. W_4 is a 3136x1024 shared weight matrix, and b_4 is the bias.

Finally, we fully connect this F5 layer to the output layer, where the number of neurons equals the number of classes using the Equation 7.14.

The results (Table 7.10) show that the recognition rate of NPW features can be significantly increased (up to 10%) by applying it on the four directional Kirsch edge images (NPW-Kirsch method). Then by combining the NPW-Kirsch features, HoG features, and Zoning method we can increase the recognition rate up to 85% [14]. In these experiments, the number of training dataset for each classes is not balanced. This condition was already clearly stated and can not be avoided in our case of IHCR development for Balinese script on palm leaf manuscripts. Some ancient characters are not frequently found in our collection of palm leaf manuscripts.

7.3.2.2 Experiment Part 2

For the second part, the training based method with a neural network and unsupervised feature learning was investigated [16] to increase the recognition rate from the previous one. We performed empirical experiments with different values of parameters for the number of hidden layers (1 or 2 layers), the number of neurons in each layer (100, 200, 300, 500), and also the learning rate (gradient step). In our opinion, a smaller number of glyph classes for the categorized recognizer does not always yield a smaller optimal number of neurons. Because for some glyph categories, the similarities between glyphs are stronger, so it needs a larger number of feature learners. We found that one single layer with 500 neurons gives the best recognition rate, and we applied this parameter setting for all recognizers. All isolated glyph recognizers were built with 500 cluster centers for the unsupervised feature learning. The recognition rates of each recognizer is shown in Table 7.11 and 7.12. Based in Table 7.11, the recognition rates of the global glyph recognizer are not too discriminative, but it shows that the single layer neural network with the initialized unsupervised feature learning (NN+UFL) generally improved the recognition rate compared to the latest study of Balinese glyph recognition with convolutional neural network (CNN) and the k-nearest neighbours (k-NN) [14]. The CNN was trained from zero on the AMADI_LontarSet dataset [15]. In this experiment we performed the glyph

Table 7.10: Recognition rate from all schemes of experiment [14]

No	Features	Feature Dim.	Classifier	Recog. Rate %
1.	Histogram Projection (Binary)	100	SVM	26.31
2.	Celled Projection (Binary)	500	SVM	49.94
3.	Celled Projection (Binary)	500	k-NN	76.16
4.	Distance Profile (Binary)	200	SVM	40.13
5.	Distance Profile (Binary)	200	k-NN	58.95
6.	Distance Profile (Skeleton)	200	SVM	36.77
7.	Crossing (Binary)	100	SVM	15.00
8.	Zoning (Binary)	205	SVM	50.65
9.	Zoning (Binary)	205	k-NN	78.54
10.	Zoning (Skeleton)	205	SVM	41.85
11.	Zoning (Grayscale)	205	SVM	52.42
12.	Zoning (Grayscale)	205	k-NN	66.13
13.	Gradient Feature (Gray)	400	SVM	60.04
14.	Gradient Feature (Gray)	400	k-NN	72.58
15.	Moment Hu (Gray)	56	SVM	33.48
16.	Moment Hu (Gray)	56	k-NN	33.48
17.	HoG (Gray)	1984	SVM	71.28
18.	HoG (Gray)	1984	k-NN	84.35
19.	NPW (Binary)	100	SVM	51.39
20.	NPW (Gray)	100	SVM	54.13
21.	Kirsch (Gray)	100	SVM	62.45
22.	HoG with Zoning (Gray)	1984	SVM	69.69
23.	HoG with Zoning (Gray)	1984	k-NN	83.50
24.	NPW-Kirsch (Gray)	400	SVM	63.57
25.	NPW-Kirsch (Gray)	400	k-NN	76.71
26.	HoG on Kirsch edge (Gray)	1984*4	k-NN	82.09
27.	HoG + NPW-Kirsch (Gray)	1984+400	k-NN	84.75
28.	Zoning + Celled Projection (Binary)	205+500	k-NN	77.70
29.	HoG + NPW-Kirsch (Gray) + Zoning (Binary)	1984+400 +205	k-NN	**85.16**
30.	CNN 1			84.31

Table 7.11: Recognition rate of the global glyph recognizer

Glyph Classes	Number of Data	CNN 1 [14]	k-NN [14]	NN	NN UFL
133	Train : 11,710, Test : 7,673	84.31	85.16	85.51	**85.63**

Table 7.12: Recognition rate of the categorized glyph recognizer

Category	Glyph Classes	Number of Data Subsets	NN	NN UFL
ASC	7	Train : 860, Test : 921	92.73	**93.16**
DESC	20	Train : 1,860, Test : 593	85.84	**88.03**
BASE	49	Train : 5,070, Test : 4,392	**87.46**	87.43
ASC-BASE	16	Train : 1,170, Test : 208	75.48	**75.96**
BASE-DESC	40	Train : 2,550, Test : 1,309	86.40	**86.63**

recognition of NN+UFL with exactly the same dataset. Table 7.12 shows that the use of NN+UFL for the categorized glyph recognizers gives generally more positive effects. Following our goal to validate the result of global recognition, the recognition rates of a categorized glyph recognizer is generally higher than the recognition rate of the global glyph recognizer. It means that if the spatial category detection of a glyph is correct, the categorized glyph recognizer will recognize the glyph better than the global glyph recognizer. We have not discredited global glyph recognition, but it depends on some conditions of the option selection rules. The final choice of glyph recognition is actually a well-structured compromise between global and categorized glyph recognition.

7.3.2.3 Experiment Part 3

In the third part, we conduct a broader evaluation of the robustness of the methods previously tested on Balinese script, to the other two different palm leaf manuscripts with Khmer and Sundanese scripts. The experimental results for the isolated character/glyph recognition task are presented in Table 7.13. For handcrafted feature with k-NN, the Khmer set with 113,206 train images and 90,669 test images will need a considerable amount of time for one-to-one k-NN comparison, so we do not think it is reasonable to use it. For CNN 1, previous work only reported results for the Balinese set. For all ICFHR competition methods, the competition was proposed only for the Balinese set, so we only have the reported results for the Balinese set.

For the second architecture of CNN, a vanilla CNN is used. The architecture of the CNN (Figure 7.29) is described as follow (this architecture has also been reported in Khmer isolated character recognition baseline in [117]. The grayscale input images of isolated characters are rescaled to 48x48 pixels in size and are normalized by applying histogram stretching. The network consists of three sets of convolution and max pooling pairs. All convolutional layers use a stride of one and are zero padded so that the output is the same size as the input. The output of each convolutional layer is activated using a ReLu function and is followed by a max pooling of 2x2 blocks. The number of feature maps (of size 5x5) used in the three consecutive convolutional layers are 8, 16, and 32 respectively. The output of the last layers is flattened, and a fully-connected layer with 1024 neurons

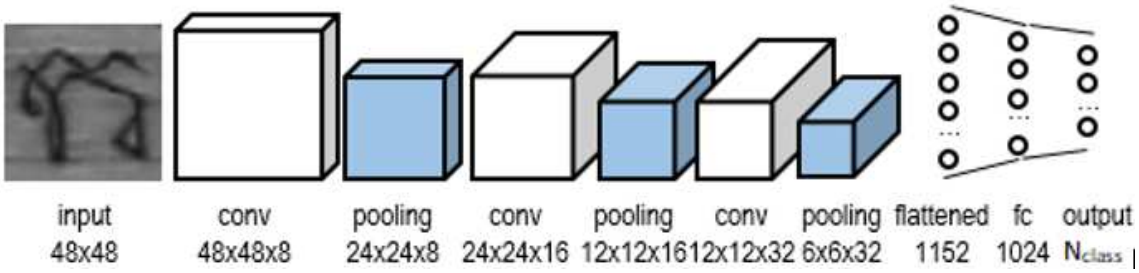


Figure 7.29: Architecture of the CNN

(also activated with ReLu) is added followed by the last output layer (softmax activation) consisting of N_class neurons, where N_class is the number of character classes. Dropout with probability p=0.5 is applied before the output layer to prevent overfitting. We trained the network using the Adam optimizer with the batch size of 100 and the learning rate of 0.0001.

According to these results, the handcrafted feature extraction combination of HoG-NPW-Kirsch-Zoning is a proper choice to produce good recognition rate for Balinese and Khmer characters/glyphs. The CNN 2 methods show also satisfactory results, but the differences in recognition rate are not too significant with the handcrafted feature combinations. The unbalance number of image samples for each character class makes the CNN methods not perform good enough. For the Sundanese dataset, the handcrafted features with NN slightly outperforms the CNN method. The UFL method increases slightly the recognition rate of pure NN method for the Khmer and Balinese datasets. In the ICFHR Competition 2016 [23], the two methods proposed by Group 1 achieved the top two recognition rates. Group 1 obtained a significant 10% higher recognition rate than two other groups. By extracting the gradient based feature for the virtual images to expand the training set, the method VMQDF from Group 1 outperforms all other methods.

7.3.3 Conclusions

Balinese scripts on palm leaf manuscripts offer a real new challenge in IHCR development. We present our experimental study on feature extraction methods for character recognition of Balinese script on palm leaf manuscripts by investigating ten feature extraction methods for IHCR. We proposed the proper and robust combination of feature extraction methods to increase the recognition rate. Our study shows that the recognition rate can be significantly increased by applying NPW features on four directional Kirsch edge images. And the use of NPW on Kirsch features in combination with HoG features and Zoning method can increase the recognition rate up to 85%, and it still slightly better than using the convolutional neural network. The VMQDF method from ICFHR G1 uses also the gradient based feature. It confirms our finding in using HoG as a good feature combination to recognize the Balinese script. To achieve a better recognition rate, the training set for VMQDF was expanded by generating the virtual images.

The problem of inadequate training data also influences glyph recognition. The unbalanced number of image samples for each character class means the CNN methods did not perform well enough in glyph recognition. The differences in the recognition rates of the CNN methods are not too significant with the handcrafted feature combinations.

Table 7.13: Experimental results for isolated character/glyph recognition task (in % recognition rate)

Methods	Balinese	Khmer	Sundanese
Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with k-NN [14]	Test = 85.16	-	Test = 72.91
Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with NN [16]	Train = 97.91 Test = 85.51	Train = 98.48 Test = 92.15	Train = 99.93 Test = **79.69**
Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with UFL + NN [16]	Train = 98.11 Test = **85.63**	Train = 99.24 Test = 92.44	Train = 99.96 Test = 79.33
CNN 1 [14]	Test = 84.31	-	-
CNN 2	Test = 85.39	Test = **93.96**	Test = 79.05
ICFHR G1 : VCMF [23]	Test = 87.44	-	-
ICFHR G1 : VMQDF [23]	Test = **88.39**	-	-
ICFHR G3 [23]	Test = 77.83	-	-
ICFHR G5 [23]	Test = 77.70	-	-

The recognition rates of the global glyph recognizer are not too discriminative, but it shows that the single layer neural network with the initialized unsupervised feature learning (NN+UFL) generally improved the recognition rate compared to Balinese glyph recognition with convolutional neural network (CNN) and the k-nearest neighbours (k-NN). The use of NN+UFL for the categorized glyph recognizers gives generally more positive effects.

7.4 Glyph Segmentation and Recognition Evaluation

To determine the combined performance between glyph segmentation and glyph recognition, we performed an experiment to evaluate the sequence of process from text line segmentation and glyph segmentation to the glyph recognition as already described from Section 6.1 to 6.5. As already mentioned in Section 5.4, 19 manuscript pages with a complete glyph segmentation and annotation are provided to test and to evaluate the glyph segmentation and recognition scheme.

7.4.1 Evaluation Metrics

The ***Segmentation Rate*** (***SR***) is calculated. It is defined as the percentage of correctly overlapped (>50%) glyph segments between the result file and the ground truth file (***Number Segments Overlapped*** = ***NSO***) over the total number of glyph segments in the result file (***Number Segments Result*** = ***NSR***). We can not calculate the ***SR*** over the total number of glyph segments in ground truth file (***Number Segments Ground Truth*** = ***NSG***), because in a ground truth file there may be double glyph segments overlapped on a certain area. The ***Segmented Recognition Rate*** (***SRR***) is then calculated. It is defined as the percentage of correctly recognized glyph segments in the result file (***Number Recognized Result*** = ***NRR***) over the ***NSO*** (Figure 7.30).

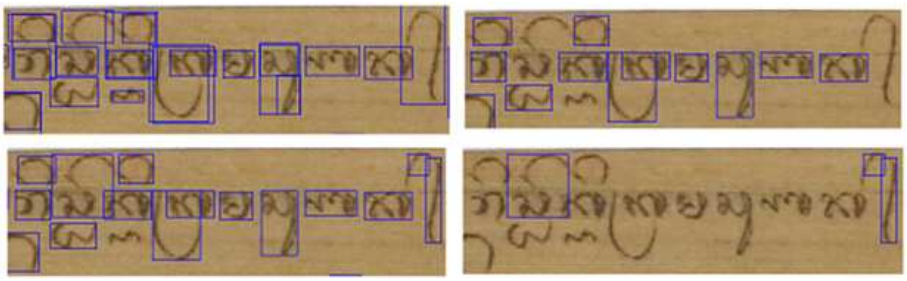


Figure 7.30: Top left: ground truth, Bottom left: glyph segments result, Top right: correctly overlapped glyph segments, Bottom right: wrong glyph segments

7.4.2 Results and Discussion

Figure 7.31 shows an example of glyph segmentation results. Figure 7.32 shows an example of glyph segmentation and recognition result. Table 7.14 shows the value of ***NSG***, ***NSR***, ***NSO***, ***SR***, ***NRR***, and ***SRR*** for the test manuscript pages. Generally, the manuscripts number 3-4 and 9-16 have high enough SR and SRR. These manuscripts come from museum collection with a better condition of preservation. Other manuscripts are the private family collection with more degradation so it is harder to segment and to recognize the glyphs. A specific remark for the manuscript number 10-11 and 17-19, these manuscripts were written in a particular column format and it makes the text line segmentation not work properly, so the ***SR*** and ***SRR*** are lower. This result will also be justified clearly in the experiments dedicated to the transliteration.

7.4.3 Conclusions

Generally, the manuscripts that come from museum collections with a better condition of preservation have high enough SR and SRR. Other manuscripts are the private family collection with more degradation so it is harder to segment and to recognize the glyphs.

7.5 Transliteration Evaluation

We evaluated the segmentation based and segmentation free transliteration method. For the segmentation based transliteration method, we evaluated first the correctness of the phonological rules of the Balinese script transliteration. For the segmentation free transliteration method, we evaluated the word transliteration and text line transliteration.

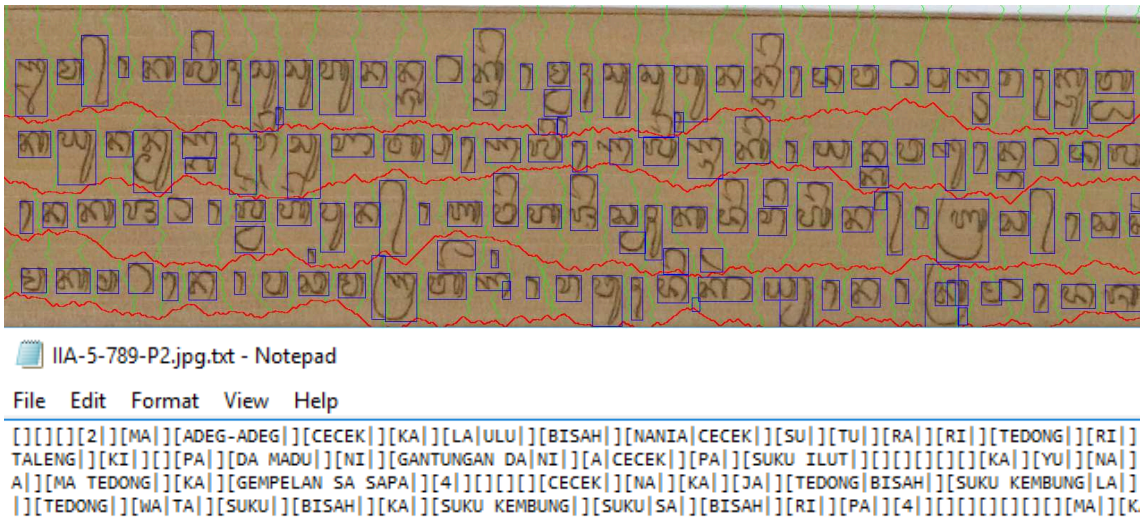
7.5.1 Evaluation Metrics

Two metrics are used to evaluate the transliteration results. The Character Error Rate (CER) is calculated to evaluate the word transliteration. The recall pattern rate (***RPR***) and the precision pattern rate (***PPR***) are calculated to evaluate the text line transliteration.

- For word transliteration. The Character Error Rate (CER) is calculated. The CER is defined by edit distance metric between ground truth and recognizer output and is computed by using the provided OCRopy function ocropus-errs[8]. This distance is defined as the ratio of insertion, deletion and substitution actions compared to the total length of the word [29].

[8]https://github.com/tmbdev/ocropy/blob/master/ocropus-errs

Figure 7.31: Evaluation of Glyph Segmentation: a) Ground Truth Glyph Segmented Image, b) Glyph Segmentation Result, c) Correct Glyph Segmentation, d) Bad Glyph Segmentation



IIA-5-789-P2.jpg.txt - Notepad

File Edit Format View Help

[][][][2|][MA|][ADEG-ADEG|][CECEK|][KA|][LA|ULU|][BISAH|][NANIA|CECEK|][SU|][TU|][RA|][RI|][TEDONG|][RI|]
TALENG|][KI|][][PA|][DA MADU|][NI|][GANTUNGAN DA|NI|][A|CECEK|][PA|][SUKU ILUT|][][][][][][KA|][YU|][NA|]
A|][MA TEDONG|][KA|][GEMPELAN SA SAPA|][4|][][][][CECEK|][NA|][KA|][JA|][TEDONG|BISAH|][SUKU KEMBUNG|LA|]
|][TEDONG|][WA|TA|][SUKU|][BISAH|][KA|][SUKU KEMBUNG|][SUKU|SA|][BISAH|][RI|][PA|][4|][][][][][][][MA|][K

Figure 7.32: Glyph Segmentation and Recognition Result

Table 7.14: Results of glyph segmentation and recognition

No	Manuscript Page	NSG	NSR	NSO	SR	NRR	SRR
1	Bangli-P41	357	216	107	49,54	62	57,94
2	Bangli-P47	326	134	79	58,96	44	55,70
3	IIA-10-1534-P8	654	517	400	77,37	336	**84,00**
4	IIIC-24-1641-P8	699	502	388	77,29	266	68,56
5	JG-01-P3	653	544	383	70,40	296	77,28
6	JG-02-P6	148	130	40	30,77	11	27,50
7	JG-02-P7	160	140	57	40,71	30	52,63
8	JG-05-P8	600	405	246	60,74	164	66,67
9	MB-AdiParwa(Purana) -5338.2-IV.a-P30	587	423	329	**77,78**	256	77,81
10	MB-AjiGriguh -5783-107.2-P11	122	180	54	30,00	40	74,07
11	MB-AjiGriguh -5783-107.2-P8	175	142	31	21,83	18	58,06
12	MB-ArjunaWiwaha -GrantangBasaII-P15	302	294	191	64,97	143	74,87
13	MB-ArjunaWiwaha -GrantangBasaII-P28	310	223	169	75,78	121	71,60
14	MB-TaruPramana-P3	614	405	288	71,11	186	64,58
15	MB-TaruPramana-P4	340	426	187	43,90	131	70,05
16	MB-TaruPramana-P6	281	310	138	44,52	75	54,35
17	WN-P5b	369	317	146	46,06	65	44,52
18	WN-P7a	186	179	20	11,17	3	15,00
19	WN-P9a	191	279	61	21,86	26	42,62

MB-AdiParwa(Purana)-5338.2-IV.a-P30.jpg_line_1
GT:73.5484
tadharMANINGATUNGGATUNGga,leGASARATADANASARWAbARANAarTAMASwastramulIADAWALA.yakaTAMADIANINGASARAPATUNGAN.NIS TASARATUNGA,legADANASARWAboJARAPAna.uTAMANINgsa
Result:56.3758
wadhewrene0,waNGATUNGA,2GASARAsADANAwwabhARANAhARWAMASumulADAWALA,,yenadAMADhIANIASARAPAwaNGANINi,jah,NISTAS ARATU2gnabhanayarwuaneJARANA,,3TAMANIsa4a

MB-AdiParwa(Purana)-5338.2-IV.a-P30.jpg_line_2
GT:72.619
RATUNGA,LEGASARASAsAJININGMATI,DINuLURASARwasekARWIJAWIJAN,NINGAsep,yekAUTAMANINGDHARmasuRATUNGA.mUAHNIStAPA MidARATekengTUNON,anruWUSKArSAWASumeNGKAhengPATAWUlan,tiNADI
Result:64.9425
dha5RATUNGA,LEGASARASAcAJINIMATI,DINiLURAyawua,rngWIJARWIJAN,NINGAupapa,,yerAUTAMANIdhmamasngRATUNGA,,,,wUAHNISAPA Mm,DHARATkeiTUNON,jnsatuaWUSKASAWASaNGKAiePATAWUpanr,tarNADI

MB-AdiParwa(Purana)-5338.2-IV.a-P30.jpg_line_3
GT:71.5232
nan.madiANINGAMIDARA,RIhUWUSNINGsawATINUNU,TikaADIN.uTAMANINGAMIDARA,brasTATAANAWUTIkangsaWA,TinADINAN.yeKATATAk ausAPARIPAMIDANGAn,muangKAUCAPATATABASA
Result:60.2564
5nrabamadhwANINGAjmwidhARA,RIUWUSuwuATINUNU,TarkADINAca,,waTAMANIAMIDharaung4stajATAANAWUTIisngWA,TunnnINAN,,pena KATATAlAUCAPARIPAMadhANGA,wiKAUCAPATATABASA

Figure 7.33: The text pattern (in red) extracted between transliterated ground truth text and transliteration result text

```
GT:76.25
MPANGMALIHagancIAN,YAnnoRAMANUTUGANGWenangDANDA,1250.YANANASAYANOraASAJATIANGAntUKANGSAYA,
ringklIANGMIWAHPANYARIKANsINALIHTUnggilwenangDANDA,GUNGArTA,250,KANGSA
Result:77.8523
MPANGMALIHegadIAN,YA,naRAMANUTUGANGWnaDANDA,1250,,0rAYANANASeYANO24ASAJATIANGAdUKANGSAYARI
nnliyaMIWAHPANYARIKANamINALIHTUgiilwanadANYA,GUNGATA,250,KA
```

Figure 7.34: The text pattern (in capital letters) extracted between transliterated ground truth text and transliteration result text

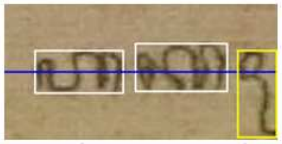
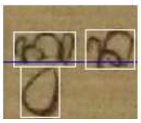
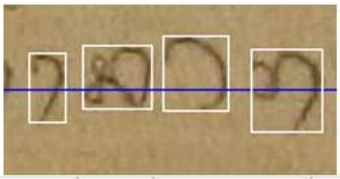
- For text line transliteration. The recall pattern rate (***RPR***) and the precision pattern rate (***PPR***) are calculated for each text line. The pattern rate (***PR***) is defined as the percentage of the same text pattern between the transliteration result text and the transliterated ground truth text. So ***RPR*** is define as ***PR*** over the length of the transliterated ground truth text, and ***PPR*** is computed as ***PR*** over the length of the transliteration result (Figure 7.33 and 7.34). The text pattern is extracted by generating the generalized suffix tree [19] between two transliterated texts (Figure 7.35 and 7.36). The minimal length of pattern text is four characters. It is defined as the minimal length of a possible transliterated Balinese word which is composed by two basic glyphs.

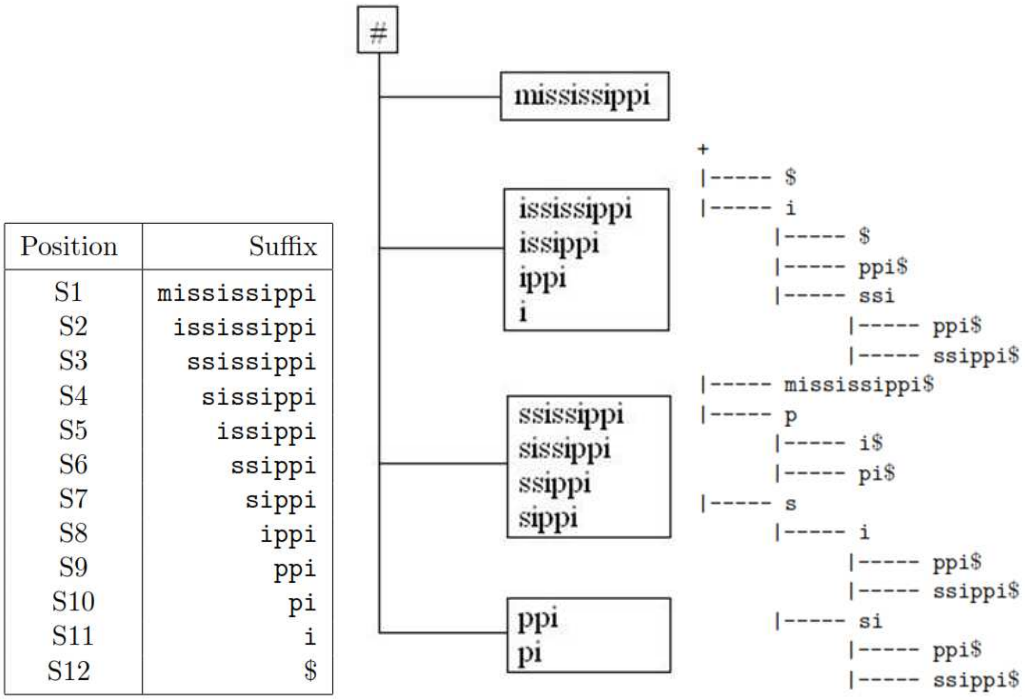
7.5.2 Segmentation Based Transliteration Evaluation: Results and Discussion

7.5.2.1 Phonological Rules Evaluation

To evaluate the correctness of phonological rules, the experiments of the transliteration process for sample word segments of Balinese palm leaf manuscript were performed for all rules (see some examples in Table 7.15). Before the transliteration process, OCR for Balinese script was performed. The evaluation of the correctness of phonological rules is started with a perfect segmented and OCRed glyphs.

Table 7.15: OCR output and phonological rules for transliteration of some sample word segments from Balinese palm leaf manuscript

OCR Output	Phonological Rules Output
 [] [] [] 'A' 'KA' 'BISAH_... [] [] []	RULE1:... RULE8:...a a RULE1:...K RULE8:...Ka RULE15:...KaH aKaH RULE32:...* aKaH* Final Output: AKAH
 [] [] 'KA' 'NA' 'SUKU KEMBUNG' []	RULE1:...K RULE2:...KW RULE12:...KWa KWa RULE1:...N RULE8:...Na KwaNa Final Output: KWANA
 [] [] [] [] 'NI' 'TALENG' 'WA' 'BISAH_... [] [] [] []	RULE17:...NI NI RULE32:...* NI* RULE1:...W RULE3:...WE RULE15:...WEH NI*WEH RULE32:...* NI*WEH* Final Output: NIWEH
 [] [] [] [] 'TALENG' 'NA' 'TEDONG' 'RA' [] [] [] []	RULE32:...* RULE1:...N RULE4:...No No RULE32:...* No* RULE1:...R RULE8:...Ra No*Ra Final Output: NORA



Position	Suffix
S1	mississippi
S2	ississippi
S3	ssissippi
S4	sissippi
S5	issippi
S6	ssippi
S7	sippi
S8	ippi
S9	ppi
S10	pi
S11	i
S12	$

Figure 7.35: Example of Suffix Tree for the string "*mississippi*" [19]

7.5.2.2 Text Line Transliteration Evaluation

We evaluated the complete process of transliteration starting from the original color palm leaf manuscript images. This evaluation is an end-to-end test by using the proposed complete scheme which was described in Chapter 6. The results are evaluated with the ground truth of the transliterated text provided by philologists. Our scheme shows a very promising result for Balinese palm leaf manuscripts transliteration and can be adapted to other types of script (Figure 7.37).

Table 7.16 shows the maximum, minimum, and average value of ***RPR*** and ***PPR*** for each manuscript collection from the test manuscript pages. The errors are cumulating from one step of the process to the other, because the previous steps are generally not perfect. The collection number 2-11,19, and 21-22 are the collection from the museum. The RPR and PPR from these collections are high enough. By extracting between 30-50% transliterated text pattern, our proposed method will likely to be usable to extract and to index some keywords on future palm leaves.

As already explained on the previous experiment, in the collection number 20 and 23, many of the pages were written in a particular column format, so the ***RPR*** and ***PPR*** are too low because the text line segmentation task did not perform in an optimal way. A particular case that we also observed from the collection number 17 is that it contains more graphics and the very damaged page border due to the bad digitization process. In this case, the segmentation task is failed.

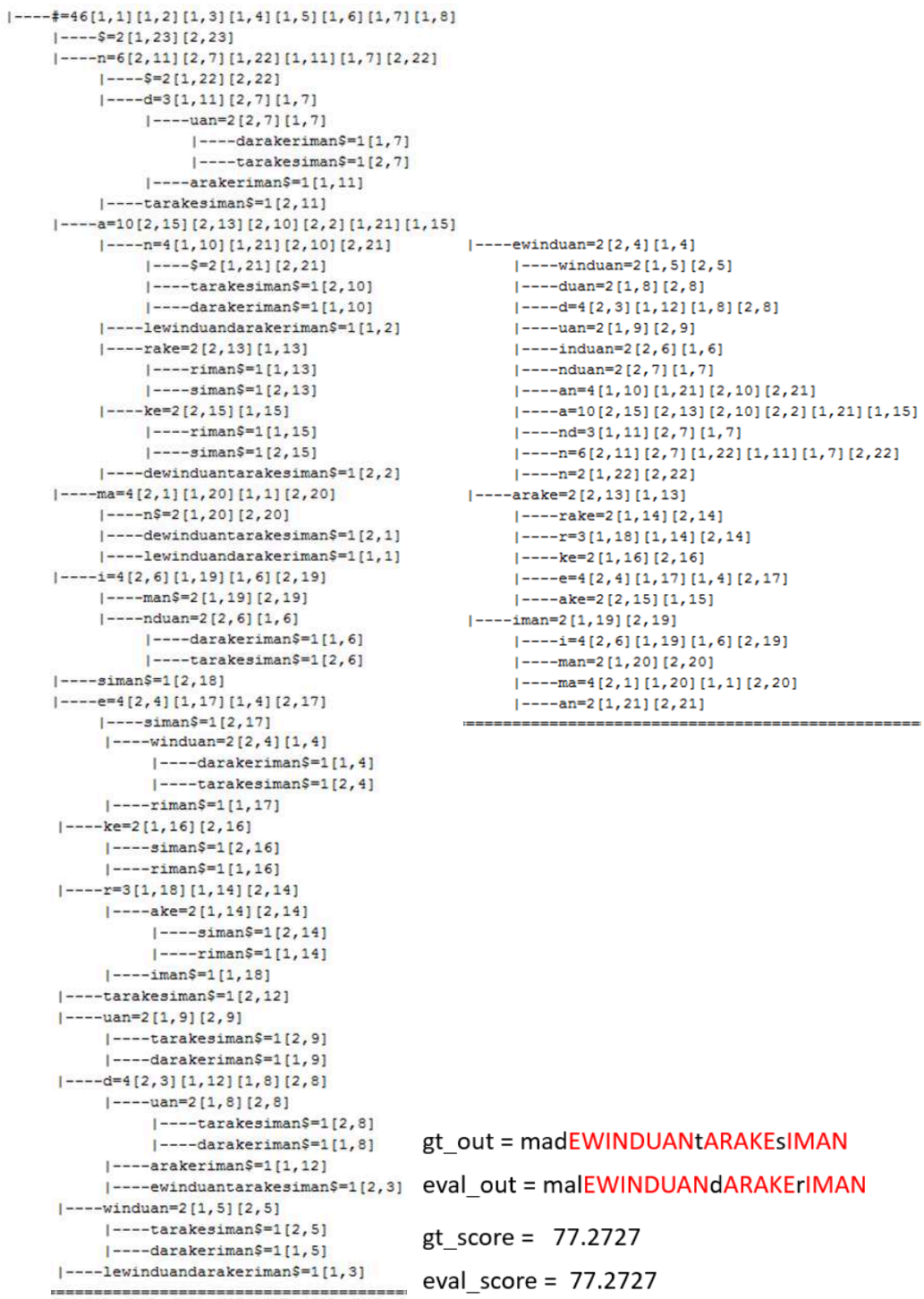
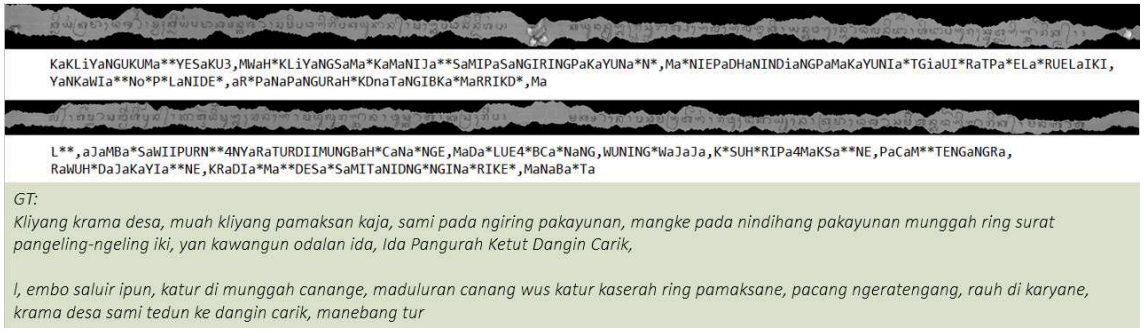


Figure 7.36: Left : Example of Generalized Suffix Tree between GT string *"madewinduan-tarakesiman"* and Evaluated string *"malewinduandarakeriman"*, Right : the Pattern Tree



GT:
Kliyang krama desa, muah kliyang pamaksan kaja, sami pada ngiring pakayunan, mangke pada nindihang pakayunan munggah ring surat pangeling-ngeling iki, yan kawangun odalan ida, Ida Pangurah Ketut Dangin Carik,

I, embo saluir ipun, katur di munggah canange, maduluran canang wus katur kaserah ring pamaksane, pacang ngeratengang, rauh di karyane, krama desa sami tedun ke dangin carik, manebang tur

Figure 7.37: Examples of transliteration results

Table 7.16: Result of the manuscript transliteration

No	Manuscript Collection	Nb Text Lines	RPR			PPR		
			Min	Max	Avg	Min	Max	Avg
1	Bangli	270	0	69,23	6,26	0	53,13	7,36
2	IIA-10-1534	30	0	73,13	47,78	0	68,16	43,73
3	IIA-5-789	30	0	76,25	50,27	0	**77,85**	48,67
4	IIB-2-180	29	0	72,63	**54,04**	0	69,61	**52,69**
5	IIIB-12-306	30	0	67,78	34,84	0	64,91	33,43
6	IIIB-42-1526	30	0	78,61	51,50	0	71,43	50,06
7	IIIB-45-2296	29	0	65,09	40,29	0	65,61	39,15
8	IIIC-19-1293	30	0	60,38	35,52	0	63,40	36,56
9	IIIC-20-1397	29	0	75,58	34,44	0	60,48	34,46
10	IIIC-23-1506	18	0	54,46	27,98	0	56,82	25,02
11	IIIC-24-1641	26	0	60,44	40,61	0	54,49	41,91
12	JG-01	62	0	56,47	23,88	0	57,72	20,70
13	JG-02	22	0	11,11	0,80	0	22,86	1,53
14	JG-03	56	0	25,83	4,91	0	23,88	7,34
15	JG-04	35	0	14,49	1,37	0	15,38	1,64
16	JG-05	17	0	27,37	4,55	0	41,94	8,85
17	JG-06	6	0	0,00	0,00	0	0,00	0,00
18	JG-07	33	0	4,94	0,15	0	12,90	0,39
19	MB-AdiParwa(Purana) -5338.2-IV.a	157	0	**83,13**	37,43	0	72,44	36,54
20	MB-AjiGriguh -5783-107.2	17	0	4,71	0,28	0	10,81	0,64
21	MB-ArjunaWiwaha -GrantangBasaII	86	0	65,81	26,83	0	68,18	32,89
22	MB-TaruPramana	149	0	70,15	17,69	0	66,67	18,99
23	WN	75	0	20,00	0,72	0	10,53	0,68
	All: 390 pages	1266						

Table 7.17: Experimental results for word recognition and transliteration task (in % error rate for test)

Methods (with OCRopy[9] framework)	Balinese	Khmer	Sundanese
BLSTM 1 (seq_depth 60, neuron size 100)	43.13	Latin text : **73.76** Khmer text : **77.88**	**75.52**
LSTM 1 (seq_depth 100, neuron size 100)	42.88	-	-
BLSTM 2 (seq_depth 100, neuron size 200)	40.54	-	-
LSTM 2 (seq_depth 100, neuron size 200)	**39.70**	-	-

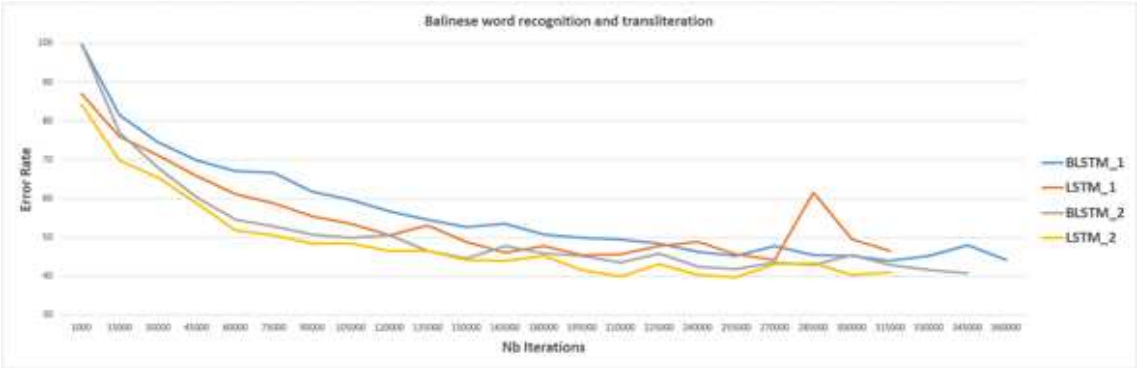


Figure 7.38: Error rate for Balinese word recognition and transliteration test set

7.5.3 Segmentation Free Transliteration Evaluation: Results and Discussion

7.5.3.1 Experiment Part 1: Word Transliteration

The experimental results for word recognition and transliteration task are presented in Table 7.17. The error rates for word recognition and transliteration test set on each training model iteration are shown in Figure 7.38-7.40. For the Khmer dataset, the transliterated text were evaluated in Khmer and Latin based alphabet (Figure 7.39). The LSTM based architecture of OCRopy seems very promising in recognizing and transliterating directly the Balinese words. For the Khmer and Sundanese dataset, the LSTM architecture seems quite difficult to learn the training data. More synthetic training data with more frequent words should be generated in order to support the training process. For the Balinese dataset, the sequence depth of 100 pixels with neuron size of 200 gives a better result for both LSTM and BLTSM architectures. Most of the Southeast Asian scripts are syllabic scripts. One character/glyph in these scripts represents a syllable, with a sequence of letters in the Latin script. In this case, word transliteration is not just word recognition with one-to-one glyph-to-letter association. This makes word transliteration more challenging than character/glyph recognition.

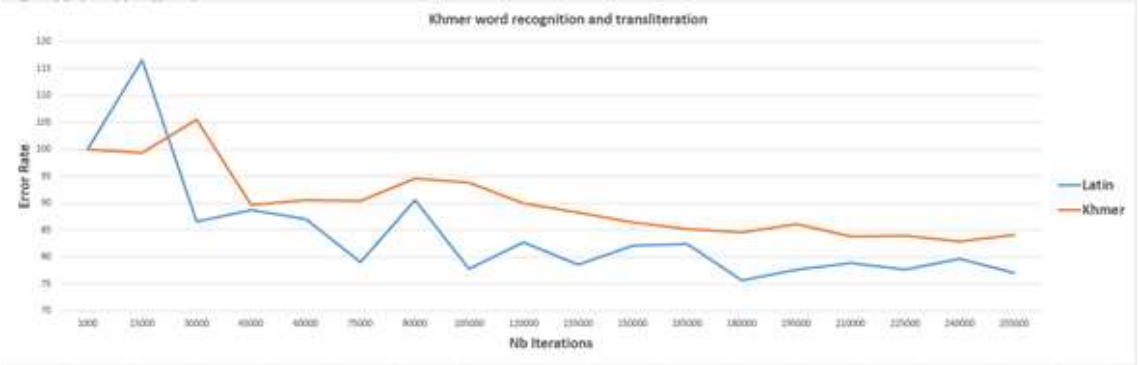


Figure 7.39: Error rate for Khmer word recognition and transliteration test set

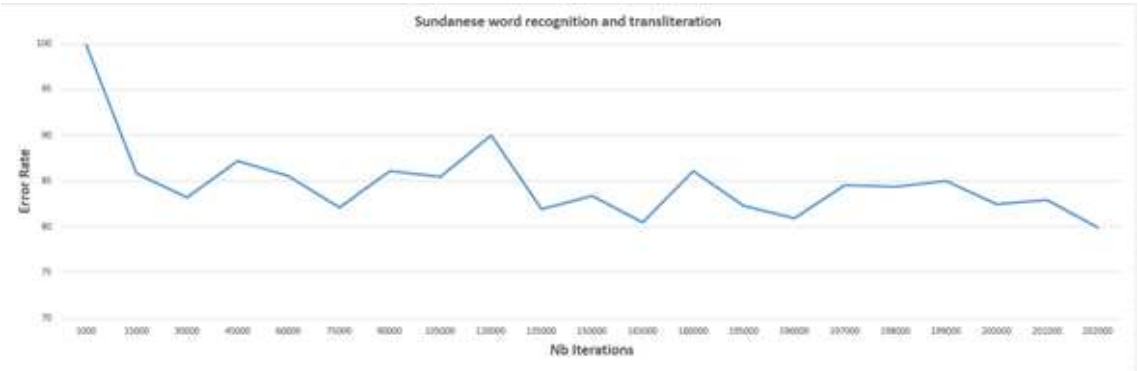


Figure 7.40: Error rate for Sundanese word recognition and transliteration test set

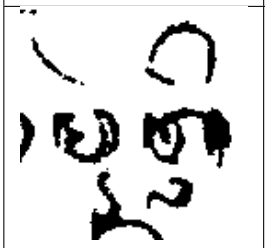
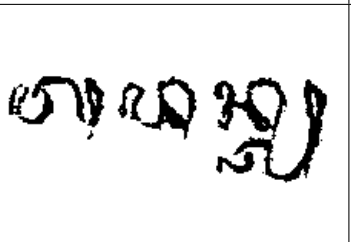
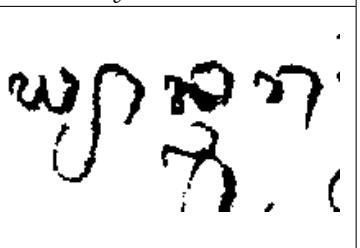
7.5.3.2 Experiment Part 2: Word and Text Line Transliteration

7.5.3.2.1 Training Schemes for the Transliteration of the Balinese Palm Leaf Manuscripts

As already explained in the Chapter 3 Sub Section 3.3.2, the challenge for the transliteration of the Balinese script comes from the fact that it is a syllabic script and that the speech sound of the syllable change related to some certain phonological rules [26]. The mapping between linguistic symbols and images of symbols is not straightforward (Figure 2.34). In addition, with a very limited training data availability, some adaptations of LSTM in transliteration training scheme need to be designed, to be analyzed and to be evaluated to ensure the robustness of the results. This work contributes in proposing and evaluating some adapted segmentation free training schemes with LSTM for the transliteration of the Balinese script into the Latin script from palm leaf manuscript images.

In this work, the training schemes at two different levels (three schemes at word level and three schemes at text line level) are proposed and are evaluated to transliterate the words and text lines of Balinese script on palm leaf manuscript. These two different levels are investigated because the absence of spaces in Balinese script makes a non trivial separation between words and text lines. The different training schemes at each level come from the availability of real annotated image samples at word level combined with the possibility of generating and using the synthetic image samples at both word and text line level. First, we briefly describe the Automatic Synthetic Handwritten Balinese Script Generator from which our synthetic image samples and the proposed training schemes

Table 7.18: Real word image samples collected from real word annotated images for Scheme W1

murti	*tadantia*	*ywantara*
		

are built.

7.5.3.2.2 The Automatic Synthetic Handwritten Balinese Script Generator

The segmentation free approach requires a high amount of training data [29]. LSTM is a supervised learning method that require a lot of ground truth training data. Annotating Balinese palm manuscripts by a human is very difficult, laborious and costly. The creation of synthetic data can decrease the tedious and costly manual works. To increase the quantity and variety of training data of real dataset from the AMADI_LontarSet, the synthetic training images have been added for this work. For this purpose, an automatic synthetic handwritten Balinese script generator has been developed prior to this work. This application generates automatically and synthetically an image of Balinese script from a Latin text to simulate the degraded handwriting sample on a Balinese palm leaf manuscript. Meaningless synthetic data training limits the advantages of segmentation-free OCR approaches [29]. For this work, to add the number of unique real words, we extracted 42,318 real sentences with 53,007 new unique real words from 41 new Latin script text corpus which were written in Balinese and Kawi language without any association with any manuscript images. The 74 real basic glyph classes of the isolated real glyph annotated images are used to render the meaningful real words and text lines into the meaningful synthetic Balinese script images.

7.5.3.2.3 Proposed Training Scheme at Word Level

Three different strategies for training at the word level are proposed.

- **Scheme W1: Training with real word image samples collected from real word annotated images**. The network is directly trained to transliterate 15,022 real word annotated image samples (see three examples in Table 7.18).

  By taking into account the condition of a limited number of available real word image samples and the possibility to generate and to use the synthetic word image samples, Schema W2 and W3 are proposed.

- **Scheme W2: Training with meaningful synthetic word image samples generated from real words (in the corpus)**. From 8,662 unique real words in the corpus, the synthetic word image samples have been generated, three image samples for each word, for a total of 25,986 image samples. The network is trained to transliterate

Table 7.19: Synthetic word image samples generated from real words (in the corpus) for Scheme W2

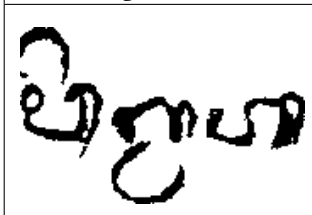
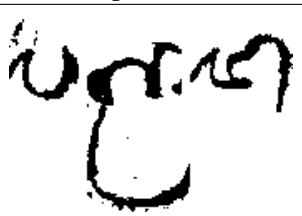
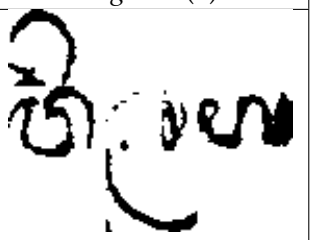
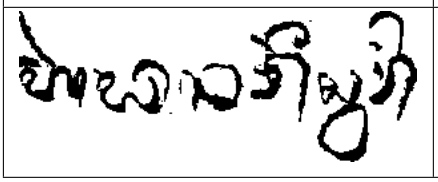
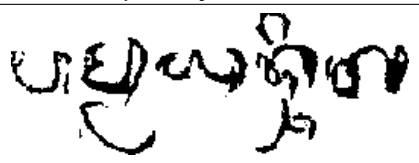
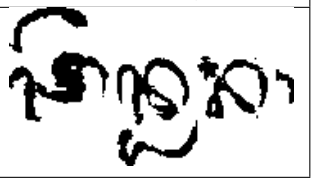
wigraha (1)	*wigraha* (2)	*wigraha* (3)
		

Table 7.20: Synthetic word image samples generated from real words (not in the corpus) for Scheme W3

arddanareswari	*pamrayashcita*	*cendana,*
		

these synthetic word image samples. Table 7.19 shows the examples of three synthetic word image samples generated from real word *wigraha*.

- **Scheme W3: Training with meaningful synthetic word image samples generated from real words from different corpus**. From 53,007 unique real words, 23,007 words have been used to generate the synthetic word image samples, one image sample for each word, for a total of 23,007 image samples. The network is trained to transliterate these synthetic word image samples (see three examples in Table 7.20)

7.5.3.2.4 Proposed Training Scheme at the Text Line Level

Three scheme for training at the text line level are proposed.

- **Scheme T3: Training with meaningful synthetic text line image samples generated from real words (not in the corpus) and with spaces between words**. The network is trained to transliterate 25,000 meaningful synthetic text line image samples, each consists of five words (with spaces between words), randomly selected from the 53,007 real words from Scheme W3 (Figure 7.41).

- **Scheme T4: Training with meaningful synthetic text line image samples generated from real words (not in the corpus) and without any spaces between words**. The same scheme as Scheme T3 but there are no spaces between words in the transliterated ground truth data training (Figure 7.42). This scheme is also evaluated because naturally there are no spaces in writing with Balinese script.

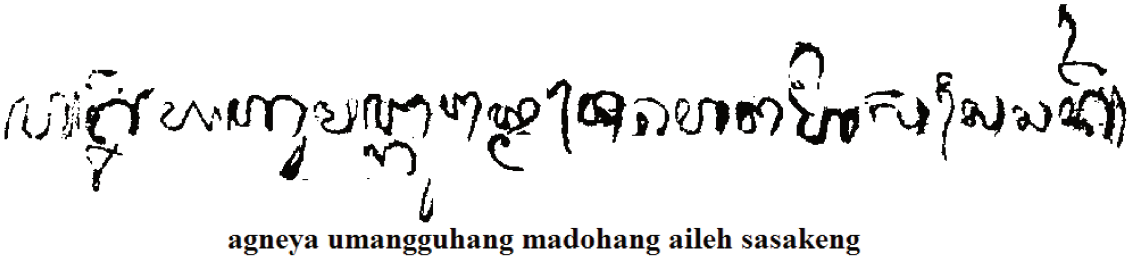


Figure 7.41: Meaningful synthetic text line image samples generated from real words (not in the corpus) and with spaces between words for Scheme T3

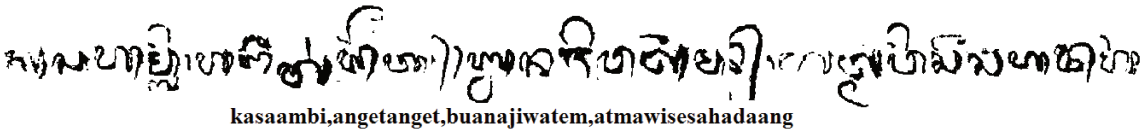


Figure 7.42: Meaningful synthetic text line image samples generated from real words (not in the corpus) and without any spaces between words for Scheme T4

- **Scheme WT (Word-Textline): Training with meaningful synthetic text line image samples generated from real sentences (not in the corpus)**. In this scheme, the pre-trained network from word level of Scheme W1 is used. The network is then trained to transliterate 25,000 meaningful synthetic text line image samples, selected and generated from the 42,318 real sentences.

7.5.3.2.5 Experimental Protocols

In this work, the OCRopy[10] framework is used to train the recognition and transliteration task for all proposed training schemes explained in Sub Section 7.5.3.2.3 with the function of ocropus-rtrain[11]. OCRopy provides the functional library of the OCR system by using RNN-LSTM architecture[12] [111]. In this work, the sequence depth of 100 pixels is used because it is more representative as the height of possible agglutination of three glyphs in vertical position. The neuron size of 200 is used, based on the best result from the previous work [24] where four variants of parameters including the unidirectional LSTM mode have already been explored and tested for Scheme W1. But the previous test and evaluation [24] were only done for the word transliteration task. The trained models are then tested to transliterate the 10,475 real word annotated images and the 1,172 real text lines images, with OCRopy function ocropus-rpred[13]. The segmentation based transliteration method (Scheme SB) [16] will also be tested and evaluated as comparison.

7.5.3.2.6 Word Transliteration Evaluation

Table 7.21 shows the error rate of word transliteration from all schemes. Scheme W1 which is trained with real word image samples gives the best result. Scheme W2 and W3 which are trained with synthetic word image samples have lower performance. Scheme W2 still gives better result than Scheme W3 because the synthetic word image samples

[10]https://github.com/tmbdev/ocropy
[11]https://github.com/tmbdev/ocropy/blob/master/ocropus-rtrain
[12]http://graal.hypotheses.org/786
[13]https://github.com/tmbdev/ocropy/blob/master/ocropus-rpred

Table 7.21: Error Rate of Word Transliteration

Scheme	W1	W2	W3	T3	T4	WT	SB
CER (%)	**39.70**	60.24	63.45	64.52	64.64	62.31	57.06

are generated from real words in the corpus. Scheme SB which is a segmentation based transliteration method gives a very promising result as the second best performance. In general, training schemes at word level perform better than training schemes at text line level. Scheme T3 and T4 which were actually trained with synthetic text lines image samples composed of synthetic word image samples give almost the same result as Scheme W3.

7.5.3.2.7 Text Line Transliteration Evaluation

Figure 7.43 and 7.44 show the average of RPR and PPR from all schemes in each collection. In general, all schemes show the same performance tendency, where they achieve better RPR and PPR for the manuscript collections from the museum (from IIA-5-789 to IIC-24-1641, and from AdiParwa to TaruPramana) compared to the manuscripts from private family collections (Bangli, JGs and WN). These latter sets correspond to the more degraded manuscripts from the private family collections. For all collections from the museum, Scheme SB with segmentation based transliteration method outperforms all segmentation free training schemes. It shows that the glyph segmentation, glyph recognition scheme and phonological rule transliteration work well for the less degraded manuscripts. Scheme WT shows that the meaningful synthetic text line image samples can be used to increase the performance of training with real word image samples in transliterating the real text line images. It also shows that training at the text line level with a pre-trained model at the word level could give a better result in word transliteration while still keeping the optimal performance for text line transliteration. The segmentation free training schemes contribute in transliterating the text lines for more degraded manuscripts, for example Scheme T4. Scheme T4 which is trained with meaningful synthetic text line image samples generated from real words (not in the corpus) and without any spaces between words gives the second best result in the overall collection. In some private family collections, Scheme T4 has a higher maximum RPR and PPR values compared to Scheme SB (Figure 7.45 and 7.46).

7.5.4 Conclusions

To build the segmentation based transliteration, a complete scheme for glyph recognition and the phonological rules for the transliteration of these manuscripts are presented in this work. The proposed scheme consists of six tasks: the text line and glyph segmentation, the detection of the spatial position for glyph category, the glyph ordering process, the global and the categorized glyph recognition, the option selection for the glyph recognition and the transliteration with phonological rules-based machine. This scheme shows a very promising result for Balinese palm leaf manuscripts transliteration and can be adapted to other type of script.

Some adapted segmentation free training schemes at two different levels (word level and text line level) with generated synthetic image samples are proposed and are eval-

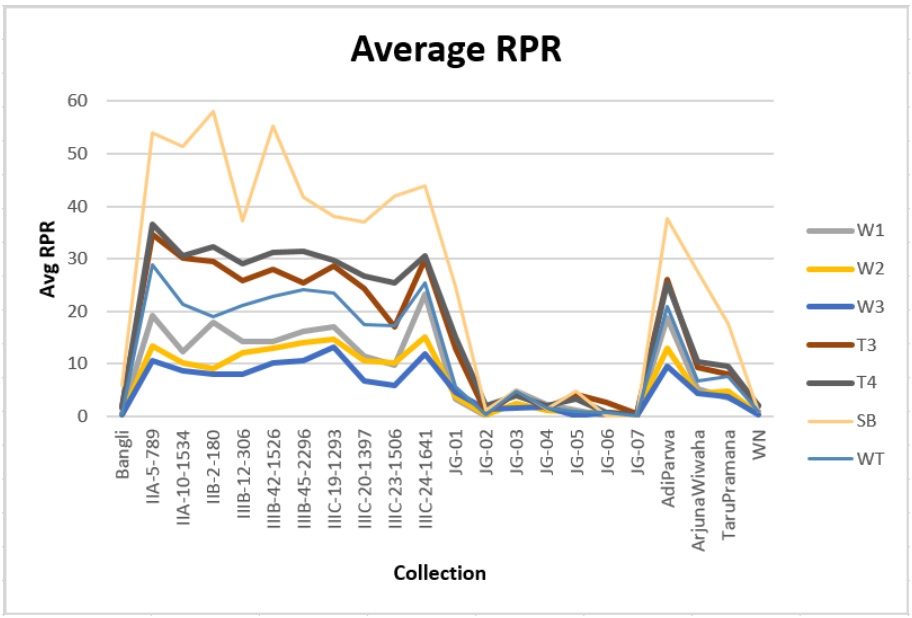


Figure 7.43: Average RPR for all collections

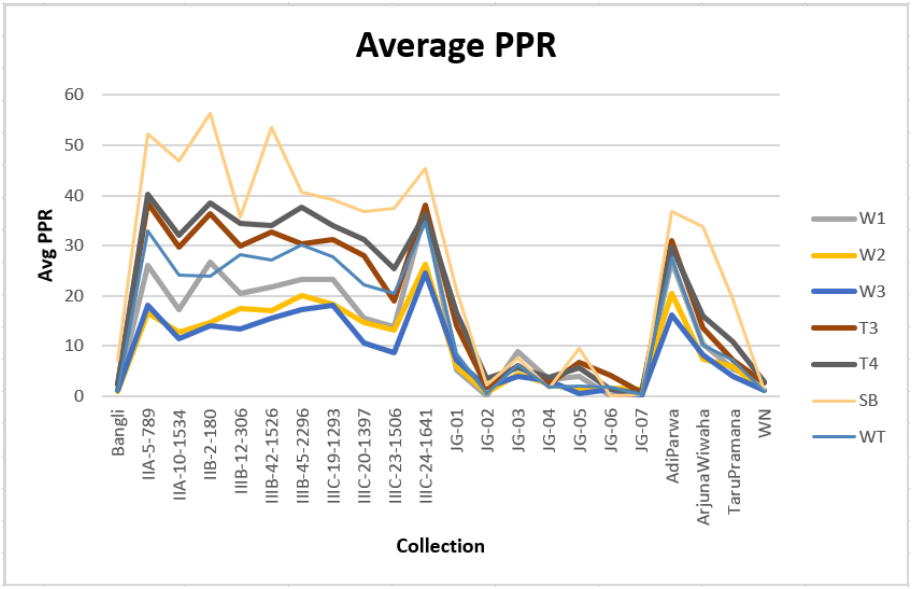


Figure 7.44: Average PPR for all collections

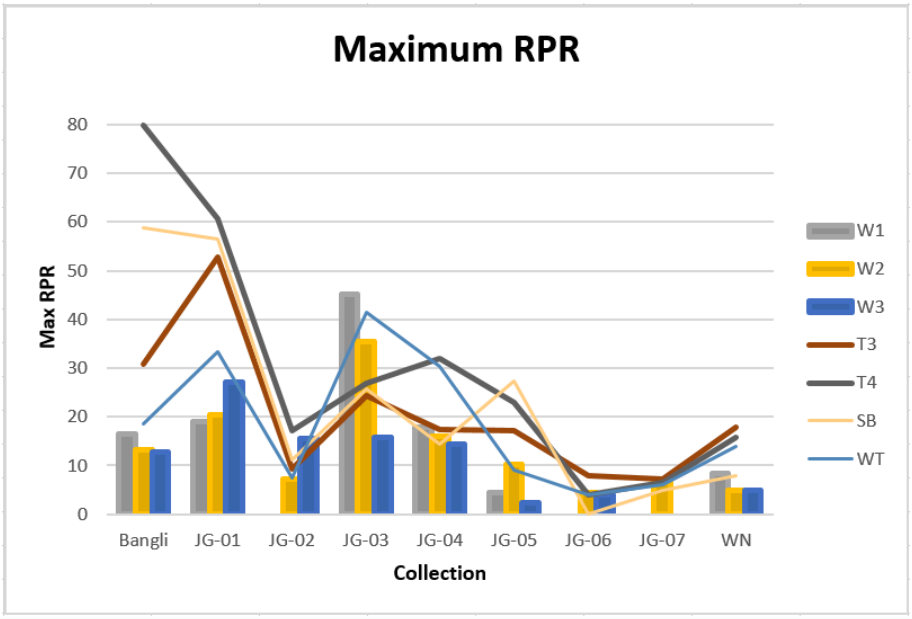


Figure 7.45: Maximum RPR only for the more degraded manuscripts from the private family collections

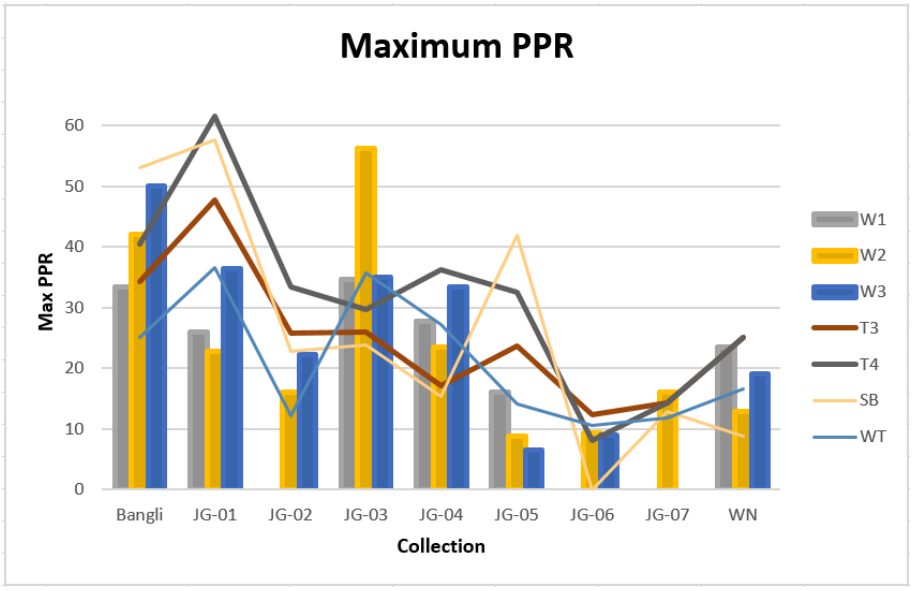


Figure 7.46: Maximum PPR only for the more degraded manuscripts from the private family collections

uated to transliterate the Balinese script into the Latin script on palm leaf manuscript images. For word transliteration, training schemes at word level perform better than training schemes at text line level. As comparison, the segmentation based transliteration method gives a very promising result as the second best performance. For text line transliteration, segmentation based transliteration method outperforms all segmentation free training schemes for the less degraded collections, while the segmentation free training schemes contribute in transliterating the text lines for more degraded manuscripts. The results show that the meaningful synthetic text line image samples can be used to increase the performance of training with real word image samples in transliterating the real text line images. Training at the text line level with a pre-trained model at the word level could give a better result in word transliteration while still keeping the optimal performances for text line transliteration. For future works, the combination of pre-trained model at the word level and text line level between real and synthetic image samples will be investigated to give more optimal performances for both word and text line transliteration.

Chapter 8

Conclusions and Future Work

This chapter finally gives some conclusions of the work presented in this dissertation by describing some limitations of the proposed system and the possible improvements for future work.

8.1 Summary

Ancient manuscripts record much important knowledge about world civilization history. In Southeast Asia, most of the ancient manuscripts are written on palm leaves. Southeast Asia is a home for many ancient manuscripts where most of those manuscripts were handwritten on the dried palm leaves with complex languages and scripts. Ancient palm leaf manuscripts store various forms of knowledge and historical records including art, religion, and local wisdom from a long time ago. Some points of interest in working with the collections of heritage document are the huge quantity of the document, the physical condition of material of the document, the very valuable cultural content of the document, and the possible access to the collection of document, the need to share and to open the document for the community and to make physical documents available to a large number of people. All four important points of interest in working with the collections of heritage document described above can be clearly seen and be identified as real challenges as well as an integral part for the background and context of the research work in this dissertation. The palm leaf manuscript collections from Bali that became the main research object in this dissertation appeared exactly at the midpoint of those four points of interest.

Following the increasing development of the preservation project for heritage documents around the world, the collection of palm leaf manuscripts in Southeast Asia finally attracted the attention of researchers in document image analysis (DIA). The digitization and indexing projects for palm leaf manuscripts were proposed. Within the scope of the heritage documents, the research work conducted for this dissertation focused on the heritage documents of the collection of palm leaf manuscripts from Indonesia, especially the palm leaf manuscripts from Bali. In some parts, the collection of palm leaf manuscripts from Sunda (West Java - Indonesia) and the Khmer palm leaf manuscripts from Cambodia were also described. This dissertation took part in exploring DIA research for palm leaf manuscript collection. This collection offers new challenges for DIA research because it uses palm leaf as the writing media and also with a language and script that have never been analyzed before.

We presented a brief overview of the existence of heritage documents in general from the beginning of history of writing materials. The interests in heritage documents from some points of view are described. We introduced the socio-cultural aspects of palm leaf manuscripts from Southeast Asia and exposed more specifically the unique characteristics of Balinese palm leaf manuscripts, the collections, the productions and the writing tools. The basic concept of the Balinese language with syllabic script is also presented.

We described and discussed the challenges for the development of DIA system of Balinese palm leaf manuscripts. There are several major challenges in working with the collection of palm leaf manuscripts. Those challenges are not only about technical challenges, but also socio-cultural challenges as a point of interest associated with the collection of Balinese palm leaf manuscripts. The first challenge is the socio-cultural aspects. The second challenge is the physical characteristic of the manuscripts, and the third challenge is the complexity of Balinese script. The first socio-cultural challenge is the difficulty in collecting sample manuscripts to create the initial corpus for the Document Image Analysis (DIA) research. The second socio-cultural challenge is the difficulty in finding the Balinese philologist to work within this project. The challenging situations

and real conditions are identified when the access to the collection and to the content of palm leaf manuscripts are limited for some following reasons: the physical condition and the fragility of the palm leaf manuscripts, the limited access to the private family collection, the linguistic difficulties, and the difficulties in searching for a certain collection in the catalog of museum or cultural agency or institution. For the technical challenges for palm leaf manuscripts in DIA system, they can be viewed in two things. The first challenge is the physical condition of the palm leaf manuscript. Most of the collections found, especially those stored in private family homes, are already in a degraded state and disrepair. Due to the specific characteristics of the physical support of the manuscript, the development of DIA methods for palm leaf manuscripts in order to extract relevant information is considered as a new research problem in handwritten document analysis. It ranges wide from the binarization process, text line segmentation, character recognition to the text transliteration tasks. These physical degradations make the binarization task, the text line segmentation and glyph segmentation task quite challenging in the DIA pipeline for Balinese palm leaf manuscripts. The second challenge is the complexity of the Balinese script. The Balinese palm leaf manuscripts with different scripts and languages surely provide some real challenges for document analysis methods, not only because the different forms of characters from the script, but also the writing style for each script differs in how to write and to join or separate a character in a text line. Those challenging situations and real conditions related to the limited access of the palm leaf manuscripts lead a multidisciplinary scientific challenges to the context of this research in socio-cultural aspect, philology, linguistic part and document image analysis.

Motivated by the contextual situations and real conditions of the palm leaf manuscript collections in Bali, this research tried to bring added value to digitized palm leaf manuscripts by developing tools to analyze, to transliterate and to index the content of palm leaf manuscripts. These systems will make palm leaf manuscripts more accessible, readable and understandable to a wider audience and, to scholars and students all over the world. This research work is not only to digitize the palm leaf manuscripts, but also to develop an automatic analysis, transliteration and indexing system for the manuscripts. This research developed a DIA system for document images of palm leaf manuscripts, that includes several image processing tasks, beginning with digitization of the document, ground truth construction, binarization, text line and glyph segmentation, ending with glyph and word recognition, transliteration and document indexing and retrieval. In this research, we constituted the first corpus and dataset of the Balinese palm leaf manuscripts for the DIA research community. We also developed the glyph recognition system and the automatic transliteration system for the Balinese palm leaf manuscripts.

The presentation of all existing methods for each task from the state-of-the-art of DIA system was given. The global overview of DIA system and the ground truth construction were firstly given. The more detailed description of existing methods for each task in the DIA pipeline was then presented. We also contributed in presenting a number of experimental evaluations and empirical benchmarkings of commonly used DIA methods and algorithms for palm leaf manuscript dataset. For binarization, we tested and compared several alternative well-known binarization algorithms, and in order to overcome the binarization problem on degraded and low quality palm leaf manuscript images, proposed a 'semi-local' concept to apply a powerful global binarization method on only precise local character area (see the methods in Sub Section 4.3.1 and 5.3.2.1, and the evaluations in Section 7.1). For text line segmentation, we investigated the performance of text line

segmentation methods by conducting comparative experimental studies on the collection of Southeast Asian palm leaf manuscript images (see the methods in Sub Section 4.3.2 and the evaluations in Section 7.2). We also proposed the use of an ellipse energy function to significantly improve the performance of the recursive tracing function from the shredding method. For isolated glyph recognition, we investigated and evaluated some most commonly used features for character recognition, proposing and evaluating the combination of features, and implementing the supporting glyph recognition for the transliteration of Balinese script (see the methods in Sub Section 4.3.3 and the evaluations in Section 7.3). For word transliteration, we evaluated the segmentation free LSTM based method for word transliteration of Southeast Asian palm leaf manuscript images (see the methods in Sub Section 4.3.4, and the evaluations in Section 7.5).

We presented the corpus and ground truth dataset of Balinese palm leaf manuscripts which are collected, constructed and used for all research works in this dissertation. The protocols design and the complete process from the manuscript digitization process until the dataset annotation process are described. The additional corpus and dataset from Khmer and Sundanese palm leaf manuscripts are also presented. We contributed in constructing the Balinese palm leaf manuscript corpus, designing the ground truth tools and protocols, and finally presenting the first Balinese palm leaf manuscript dataset for the DIA researchs. In detail, we contributed in:

- collecting and digitizing the Balinese palm leaf manuscripts corpus (see Section 5.1 and 5.2)
- analyzing and proposing the need for a specific scheme for the construction of the ground truth of binarized images (see Sub Section 5.3.2)
- analyzing the human intervention subjectivity on the construction of ground truth binarized image and we measuring quantitatively the ground truth variability of palm leaf manuscript images (see Sub Section 7.1.2.2 and 7.1.2.3)
- designing the overall scheme of ground truth construction and annotation protocols for palm leaf manuscript images and presenting the AMADI_LontarSet, the first handwritten Balinese palm leaf manuscript dataset (see Section 5.3),
- organizing the competition on the document image analysis tasks for Balinese Palm Leaf Manuscripts and Southeast Asian Palm Leaf Manuscripts for a wider DIA research communities, in the 15th and the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016 and 2018). The datasets built for the competitions are now publicly available for scientific use.

We presented the scheme of DIA for Balinese palm leaf manuscripts. We described the knowledge representation and phonological rules which are built for the transliteration engine of Balinese script. The segmentation-based scheme which consists of a complete scheme of DIA and the segmentation-free scheme for the transliteration system are presented. We contributed in developing a segmentation-based glyph recognition and transliteration scheme of Balinese palm leaf manuscripts by proposing an implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscript. A rule-based engine for performing transliterations was proposed. The phonological rules were built and were formally defined based on the glyph recognition output. A rule-based engine for performing transliterations was proposed. This model is based on phonetics which is based on traditional linguistic study of Balinese transliteration. Detail description of this proposition was given

in Section 6.6. We presented a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts. The scheme consists of six tasks: the text line and glyph segmentation, the glyph ordering process, the detection of the spatial position for glyph category, the global and categorized glyph recognition, the option selection for glyph recognition and the transliteration with phonological rules-based machine. Detailed description of each task in this scheme was given in Chapter 6. We adapted a segmentation-free LSTM-based transliteration system of Balinese palm leaf manuscripts by developing the automatic synthetic handwritten Balinese script generator. This application generates automatically and synthetically an image of Balinese script from a Latin text to simulate the degraded handwriting sample on a Balinese palm leaf manuscript. We proposed and evaluated some adapted segmentation free training schemes for the transliteration of the Balinese script into the Latin script from palm leaf manuscript images. The generated synthetic dataset and the training schemes at two different levels (word level and text line level) were proposed. A detailed description of this scheme was given in Section 7.5.3.2.1.

We summarized all experimental evaluations which have been done in this research work and discussed the results for each task and step in DIA system for Balinese palm leaf manuscripts. A comprehensive experimental test of the principal tasks in a DIA system, starting with binarization, text line segmentation, and isolated character/glyph recognition, and continuing on to word recognition and transliteration for a new collection of palm leaf manuscripts from Southeast Asia, was presented. The results from all experiments provided the latest findings and a quantitative benchmark of palm leaf manuscripts analysis for researchers in the DIA community.

With the special characteristics and challenges possessed by the palm leaf manuscript collections, it requires a thorough adaptation of the application of the DIA system. Some specific adjustments need to be applied to the DIA methods for other types of documents. The adaptation of DIA for palm leaf manuscripts is not unique and is not universal for all types of problems from different collections. However, among the DIA system's non-unique solutions, one specific solution can still be designed to deliver the most optimal DIA system performance while still taking into account the conditions of that collection. The solution of a problem in the DIA project does not often consist of a single processing task, but it consists of a sequence of processing tasks from several DIA tasks to be performed in a particular pipeline. Optimization of all tasks in this pipeline depends on the type of application to be built and also depends on the condition and characteristics of the document corpus as the object of interest. For each application, the DIA pipeline task should be customized.

Based on our experiments, the average time needed to transliterate one page of palm leaf manuscript is around 2-3 minutes. This is due to the complete sequence of process that must be done starting from the process of segmentation, recognition and transliteration. Each process in the pipeline affects the final performance of the system. Although the processing time is still not very fast, but what is essential here is the first important role of our transliteration machines to be able to substitute the difficulty of finding a Balinese philologist. For the less degraded palm leaf manuscripts, the results show that by extracting between 30-50% transliterated text pattern, our proposed method will likely to be usable to extract and to index some keywords on future palm leaves. In the end, a transliteration application for palm leaf manuscripts with a simple interface, without

having to show the detail pipeline of process, must be developed to be used by the general public.

8.2 Limitations of the proposed system and possible improvements

In this research, some limitations of the proposed system and possible improvements are identified as follow.

- Incomplete glyph set.
  - As already described in this dissertation, some glyph classes of Balinese script were rarely used and were not even found in the manuscript corpus. From 156 glyphs of Balinese script complete set, only 74 annotated basic glyphs were found in the dataset of the AMADI_LontarSet. It means that the proposed system can only recognize and transliterate those limited number of glyphs. With these 74 basic glyphs and some additional compound glyphs (59 compound glyphs), it seems already sufficient to build the text recognition and transliteration. But, a bigger corpus of manuscripts should still be built in the near future. More complete glyph set will facilitate the development of better transliteration system for Balinese script.
  - Having the first DIA system as the result of this research work, it opens a new possibility to propose a bigger dataset construction based on semi automatic approach. The first proposed system from this work can initialize the preliminary analysis for the manuscripts, and then the manual procedure can be placed for further validation process. By repeating this semi-automatic approach many times for each new manuscript collection, the system can be ameliorated to get better performances. With this semi automatic procedure, the crowd sourcing method can also be implemented to accelerate the dataset construction process.
  - On the other part, for the existing glyph set, a consortium of Balinese philologists is expected to solve the problem of inconsistency in alphabet-glyph mapping between Balinese and Latin script. It can be done by proposing a standard correction for all transliterated text that have been produced for some manuscript collections. This effort will help in reducing the problem of allographs between glyphs. This condition is expected to increase the performance of the segmentation-free transliteration method which has suffered from the problem of inconsistency and allographs.
  - Another solution is to build a closed set of vocabulary from Balinese, Kawi, or even some certain Sanskrit words which normally use those rarely found special glyph set. This closed set vocabulary has to be collected manually, but the sample images can be generated synthetically with only small number of glyph samples.
- Post-transliterated correction.
  - In this system, there is no post-transliterated correction step yet. Text pattern analysis for the resulted transliterated text can be one promising possible solution to be investigated in order to correct the transliteration errors and to improve the performances.

- Palm leaf manuscripts with graphical content and special or unusual writing format.
  - The proposed DIA system in this work did not apply any special method for palm leaf manuscripts with graphical content and special or unusual writing format, e.g. writing format in column or table. Palm leaf manuscripts with graphics / images are frequently found in Bali. A graphics spotting method can be very interesting to be applied and evaluated for those kinds of manuscripts.

Appendices

Appendix A

List of Communications and Publications

This dissertation has led to the following communications and publications:

- Journal papers
    1. M.W.A. Kesiman, D. Valy, J.-C. Burie, E. Paulus, I.M.G. Sunarya, S. Hadi, K.H. Sok, J.-M. Ogier, **Southeast Asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges**, J. Electron. Imaging. 26 (2016) 11011. doi:10.1117/1.JEI.26.1.011011. [25]
    2. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Knowledge Representation and Phonological Rules for the Automatic Transliteration of Balinese Script on Palm Leaf Manuscript**, Computaciòn y Sistemas, 21(4), January 2018. [26]
    3. M.W.A. Kesiman; D. Valy; J.-C. Burie; E. Paulus; Suryani, M.; S. Hadi; Verleysen, M.; Chhun, S.; J.-M. Ogier **Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia**. J. Imaging 2018, 4, 43. Special Issue Document Image Processing. [24]
- Book chapters
    1. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, G.N.M.A. Wibawantara, and I.M.G. Sunarya, **Book Chapter 9. Historical Handwritten Document Analysis of Southeast Asian Palm Leaf Manuscripts**, in B.L.D. Bezerra, ed., Handwriting: recognition, development and analysis, Nova Science Publishers, Inc, Hauppauge, New York, 2017.
- International conference papers
    1. M.W.A. Kesiman, S. Prum, J.-C. Burie, J.-M. Ogier, **An Initial Study On The Construction Of Ground Truth Binarized Images Of Ancient Palm Leaf Manuscripts**, in: 13th Int. Conf. Doc. Anal. Recognit. ICDAR, Nancy, France, 2015. [1]
    2. M.W.A. Kesiman, S. Prum, I.M.G. Sunarya, J.-C. Burie, J.-M. Ogier, **An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts**, in: 5th Int. Conf. Image Process. Theory Tools Appl. IPTA 2015, Orleans, France, 2015: pp. 229–233. [18]

3. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, **A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript**, in: 15th Int. Conf. Front. Handwrit. Recognit. 2016, Shenzhen, China, n.d.: pp. 325–330. doi:10.1109/ ICFHR.2016.63. [11]
4. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, G.N.M.A. Wibawantara, I.M.G. Sunarya, **AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset**, in: 15th Int. Conf. Front. Handwrit. Recognit. 2016, Shenzhen, China, 2016: pp. 168–172. doi:10.1109/ICFHR.2016.39. [15]
5. J.-C. Burie, M. Coustaty, S. Hadi, M.W.A. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I.M.G. Sunarya, D. Valy, **ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts**, in: 15th Int. Conf. Front. Handwrit. Recognit. 2016, Shenzhen, China, 2016: pp. 596–601. doi:10.1109/ICFHR.2016.107. [23]
6. M.W.A. Kesiman, S. Prum, J.-C. Burie, J.-M. Ogier, **Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images**, in: 23rd Int. Conf. Pattern Recognit., Cancun, Mexico, 2016. [14]
7. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Knowledge Representation and Phonological Rules for the Automatic Transliteration of Balinese Script on Palm Leaf Manuscript**, in: CICLing, Budapest, Hungary, 2017 [26]
8. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, **A Complete Scheme Of Spatially Categorized Glyph Recognition For The Transliteration Of Balinese Palm Leaf Manuscripts**, in: 14th Int. Conf. Doc. Anal. Recognit. ICDAR, Kyoto, Japan, 2017 [16]

- International communications at scientific congresses without proceedings

1. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Image Document Analysis of Ancient Palm Leaf Manuscripts**, Presented on The 3rd International Document Image Processing (IDIPS) Summer School, Poros, Kefalonia Island, Greece, June 21-26, 2015
2. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts**, Presented on manuSciences15 Franco-German Summer School on Manuscripts, Frauenwörth Island, Chiemsee, Germany, September 6-12, 2015
3. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Automatic Transcription and Indexation of Balinese Palm Leaf Manuscripts: Character and Text Recognition of Balinese Script**, Presented on 2nd PARSEME Training School, La Rochelle, France, 27 June - 1 July, 2016
4. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Image Document Analysis of Ancient Palm Leaf Manuscripts**, Presented on Doctoral Consortium IAPR 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France (Relocated from Tunisia, Africa), August 23-26, 2015
5. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Document Image Analysis of Balinese Palm Leaf Manuscripts**, Presented on Doctoral Consortium IAPR 14th International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, November 9-15, 2017

6. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, **Document Image Analysis of Palm Leaf Manuscripts**, Presented on 2nd International Conference on Innovative Research Accross Disciplines (ICIRAD), Bali, Indonesia, 26 August 2017.

- National and regional communications at scientific congresses without proceedings

1. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **L'analyse d'image de document des manuscrits sur feuilles de palmier**, Rencontre Tête Chercheuses, Muséum d'Histoire Naturelle, La Rochelle, France, January 17, 2016
2. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Analyse de manuscrits sur feuilles de palmier numérisés de l'Asie du Sud-Est**, Rencontres de l'AFRASE (L'Association française pour la recherche sur l'Asie du Sud-Est) 2016: Corpus et langues d'Asie du Sud-Est : traitement et exploitation numériques, Maison de l'Asie, L'Ecole française d'Extrême-Orient (EFEO), Paris, France, 13 December 2016. Published in La Lettre de l'AFRASE No. 93-94, Summer 2017
3. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Analyse de manuscrits sur feuilles de palmier numérisés de l'Asie du Sud-Est : de la reconnaissance de glyphes à la translitération automatique**, Journee d'etude TALADOC, Nantes, France, 2 June 2017
4. M.W.A. Kesiman, J.-C. Burie, J.-M. Ogier, P. Grangé, **Analyse de manuscrits sur feuilles de palmier numérisés de l'Asie du Sud-Est**, Remise de Prix Mahar Schutzenberger par l'AFIDES (L'Association Franco-Indonésienne pour le Développement des Sciences), Ambassade d'Indonésie, Paris, France, 3 July 2017

Appendix B

Unicode Table for Balinese Script

The character code tables and list of character names for Balinese script in The Unicode Standard, Version 10.0.

Balinese

Range: 1B00–1B7F

This file contains an excerpt from the character code tables and list of character names for
The Unicode Standard, Version 10.0

This file may be changed at any time without notice to reflect errata or other updates to the Unicode Standard.
See http://www.unicode.org/errata/ for an up-to-date list of errata.

See http://www.unicode.org/charts/ for access to a complete list of the latest character code charts.
See http://www.unicode.org/charts/PDF/Unicode-10.0/ for charts showing only the characters added in Unicode 10.0.
See http://www.unicode.org/Public/10.0.0/charts/ for a complete archived file of character code charts for Unicode 10.0.

The Unicode Standard 10.0, Copyright © 1991-2017 Unicode, Inc. All rights reserved.

Various signs

1B00 ᬀ BALINESE SIGN ULU RICEM
= ardhacandra

1B01 ᬁ BALINESE SIGN ULU CANDRA
= candrabindu

1B02 ᬂ BALINESE SIGN CECEK
= anusvara

1B03 ᬃ BALINESE SIGN SURANG
= repha

1B04 ᬄ BALINESE SIGN BISAH
= visarga

Independent vowels

1B05 ᬅ BALINESE LETTER AKARA
= a

1B06 ᬆ BALINESE LETTER AKARA TEDUNG
= aa
≡ 1B05 ᬅ 1B35 ᬵ

1B07 ᬇ BALINESE LETTER IKARA
= i

1B08 ᬈ BALINESE LETTER IKARA TEDUNG
= ii
≡ 1B07 ᬇ 1B35 ᬵ

1B09 ᬉ BALINESE LETTER UKARA
= u

1B0A ᬊ BALINESE LETTER UKARA TEDUNG
= uu
≡ 1B09 ᬉ 1B35 ᬵ

1B0B ᬋ BALINESE LETTER RA REPA
= vocalic r

1B0C ᬌ BALINESE LETTER RA REPA TEDUNG
= vocalic rr
≡ 1B0B ᬋ 1B35 ᬵ

1B0D ᬍ BALINESE LETTER LA LENGA
= vocalic l

1B0E ᬎ BALINESE LETTER LA LENGA TEDUNG
= vocalic ll
≡ 1B0D ᬍ 1B35 ᬵ

1B0F ᬏ BALINESE LETTER EKARA
= e

1B10 ᬐ BALINESE LETTER AIKARA
= ai

1B11 ᬑ BALINESE LETTER OKARA
= o

1B12 ᬒ BALINESE LETTER OKARA TEDUNG
= au
≡ 1B11 ᬑ 1B35 ᬵ

Consonants

1B13 ᬓ BALINESE LETTER KA

1B14 ᬔ BALINESE LETTER KA MAHAPRANA
= kha

1B15 ᬕ BALINESE LETTER GA

1B16 ᬖ BALINESE LETTER GA GORA
= gha

1B17 ᬗ BALINESE LETTER NGA

1B18 ᬘ BALINESE LETTER CA

1B19 ᬙ BALINESE LETTER CA LACA
= cha

1B1A ᬚ BALINESE LETTER JA

1B1B ᬛ BALINESE LETTER JA JERA
= jha

1B1C ᬜ BALINESE LETTER NYA

1B1D ᬝ BALINESE LETTER TA LATIK
= tta

1B1E ᬞ BALINESE LETTER TA MURDA MAHAPRANA
= ttha

1B1F ᬟ BALINESE LETTER DA MURDA ALPAPRANA
= dda

1B20 ᬠ BALINESE LETTER DA MURDA MAHAPRANA
= ddha

1B21 ᬡ BALINESE LETTER NA RAMBAT
= nna

1B22 ᬢ BALINESE LETTER TA

1B23 ᬣ BALINESE LETTER TA TAWA
= tha

1B24 ᬤ BALINESE LETTER DA

1B25 ᬥ BALINESE LETTER DA MADU
= dha

1B26 ᬦ BALINESE LETTER NA

1B27 ᬧ BALINESE LETTER PA

1B28 ᬨ BALINESE LETTER PA KAPAL
= pha

1B29 ᬩ BALINESE LETTER BA

1B2A ᬪ BALINESE LETTER BA KEMBANG
= bha

1B2B ᬫ BALINESE LETTER MA

1B2C ᬬ BALINESE LETTER YA

1B2D ᬭ BALINESE LETTER RA

1B2E ᬮ BALINESE LETTER LA

1B2F ᬯ BALINESE LETTER WA

1B30 ᬰ BALINESE LETTER SA SAGA
= sha

1B31 ᬱ BALINESE LETTER SA SAPA
= ssa

1B32 ᬲ BALINESE LETTER SA

1B33 ᬳ BALINESE LETTER HA

Sign

1B34 ᬴ BALINESE SIGN REREKAN
= nukta

Dependent vowel signs

1B35 ᬵ BALINESE VOWEL SIGN TEDUNG
= aa

1B36 ᬶ BALINESE VOWEL SIGN ULU
= i

1B37 ᬷ BALINESE VOWEL SIGN ULU SARI
= ii

1B38 ᬸ BALINESE VOWEL SIGN SUKU
= u

1B39 ᬹ BALINESE VOWEL SIGN SUKU ILUT
= uu

1B3A ᬺ BALINESE VOWEL SIGN RA REPA
= vocalic r

1B3B ᬻ BALINESE VOWEL SIGN RA REPA TEDUNG
= vocalic rr
≡ 1B3A ᬺ 1B35 ᬵ

1B3C ᬼ BALINESE VOWEL SIGN LA LENGA
= vocalic l

1B3D ᬽ BALINESE VOWEL SIGN LA LENGA TEDUNG
= vocalic ll
≡ 1B3C ᬼ 1B35 ᬵ

1B3E ᬾ BALINESE VOWEL SIGN TALING
= e

1B3F ᬿ BALINESE VOWEL SIGN TALING REPA
= ai

1B40 ᭀ BALINESE VOWEL SIGN TALING TEDUNG
= o
≡ 1B3E ᬾ 1B35 ᬵ

1B41 ᬄ BALINESE VOWEL SIGN TALING REPA TEDUNG
= au
≡ 1B3F ᬾ 1B35 ᬵ
1B42 ᬅ BALINESE VOWEL SIGN PEPET
= ae
1B43 ᬆ BALINESE VOWEL SIGN PEPET TEDUNG
= oe
≡ 1B42 ᬅ 1B35 ᬵ

Sign

1B44 ᭄ BALINESE ADEG ADEG
= virama

Additional consonants

1B45 ᭅ BALINESE LETTER KAF SASAK
1B46 ᭆ BALINESE LETTER KHOT SASAK
1B47 ᭇ BALINESE LETTER TZIR SASAK
1B48 ᭈ BALINESE LETTER EF SASAK
1B49 ᭉ BALINESE LETTER VE SASAK
1B4A ᭊ BALINESE LETTER ZAL SASAK
1B4B ᭋ BALINESE LETTER ASYURA SASAK

Digits

1B50 ᭐ BALINESE DIGIT ZERO
1B51 ᭑ BALINESE DIGIT ONE
1B52 ᭒ BALINESE DIGIT TWO
1B53 ᭓ BALINESE DIGIT THREE
1B54 ᭔ BALINESE DIGIT FOUR
1B55 ᭕ BALINESE DIGIT FIVE
1B56 ᭖ BALINESE DIGIT SIX
1B57 ᭗ BALINESE DIGIT SEVEN
1B58 ᭘ BALINESE DIGIT EIGHT
1B59 ᭙ BALINESE DIGIT NINE

Punctuation

1B5A ᭚ BALINESE PANTI
= section
1B5B ᭛ BALINESE PAMADA
= honorific section
1B5C ᭜ BALINESE WINDU
= punctuation ring
1B5D ᭝ BALINESE CARIK PAMUNGKAH
= colon
1B5E ᭞ BALINESE CARIK SIKI
= danda
→ 0964 । devanagari danda
1B5F ᭟ BALINESE CARIK PAREREN
= double danda
→ 0965 ॥ devanagari double danda
1B60 ᭠ BALINESE PAMENENG
= line-breaking hyphen

Musical symbols for notes

1B61 ᭡ BALINESE MUSICAL SYMBOL DONG
1B62 ᭢ BALINESE MUSICAL SYMBOL DENG
1B63 ᭣ BALINESE MUSICAL SYMBOL DUNG
1B64 ᭤ BALINESE MUSICAL SYMBOL DANG
1B65 ᭥ BALINESE MUSICAL SYMBOL DANG SURANG
1B66 ᭦ BALINESE MUSICAL SYMBOL DING
1B67 ᭧ BALINESE MUSICAL SYMBOL DAENG
1B68 ᭨ BALINESE MUSICAL SYMBOL DEUNG
1B69 ᭩ BALINESE MUSICAL SYMBOL DAING
1B6A ᭪ BALINESE MUSICAL SYMBOL DANG GEDE

Diacritical marks for musical symbols

1B6B ᭫ BALINESE MUSICAL SYMBOL COMBINING TEGEH
1B6C ᭬ BALINESE MUSICAL SYMBOL COMBINING ENDEP
1B6D ᭭ BALINESE MUSICAL SYMBOL COMBINING KEMPUL
1B6E ᭮ BALINESE MUSICAL SYMBOL COMBINING KEMPLI
1B6F ᭯ BALINESE MUSICAL SYMBOL COMBINING JEGOGAN
1B70 ᭰ BALINESE MUSICAL SYMBOL COMBINING KEMPUL WITH JEGOGAN
1B71 ᭱ BALINESE MUSICAL SYMBOL COMBINING KEMPLI WITH JEGOGAN
1B72 ᭲ BALINESE MUSICAL SYMBOL COMBINING BENDE
1B73 ᭳ BALINESE MUSICAL SYMBOL COMBINING GONG

Musical symbols

1B74 ᭴ BALINESE MUSICAL SYMBOL RIGHT-HAND OPEN DUG
1B75 ᭵ BALINESE MUSICAL SYMBOL RIGHT-HAND OPEN DAG
1B76 ᭶ BALINESE MUSICAL SYMBOL RIGHT-HAND CLOSED TUK
1B77 ᭷ BALINESE MUSICAL SYMBOL RIGHT-HAND CLOSED TAK
1B78 ᭸ BALINESE MUSICAL SYMBOL LEFT-HAND OPEN PANG
1B79 ᭹ BALINESE MUSICAL SYMBOL LEFT-HAND OPEN PUNG
1B7A ᭺ BALINESE MUSICAL SYMBOL LEFT-HAND CLOSED PLAK
1B7B ᭻ BALINESE MUSICAL SYMBOL LEFT-HAND CLOSED PLUK
1B7C ᭼ BALINESE MUSICAL SYMBOL LEFT-HAND OPEN PING

Appendix C

Complete Glyph Set of Balinese Script

Table C.1: ”Basic” Consonants in Balinese Script

No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Ha			A	Yes
2	Na / Na Kojong			NA	Yes
3	Ca / Ca Murca			CA	Yes
4	Ra			RA	Yes
5	Ka			KA	Yes
6	Da			DA	Yes
7	Ta			TA	Yes
8	Sa / Sa Danti			SA	Yes
9	Wa			WA	Yes
10	La			LA	Yes
11	Ma			MA	Yes
12	Ga			GA	Yes
13	Ba			BA	Yes
14	Nga			NGA	Yes
15	Pa			PA	Yes
16	Ja			JA	Yes
17	Ya			YA	Yes
18	Nya			NYA	Yes

Table C.2: Conjunct Forms of ”Basic” Consonants in Balinese Script

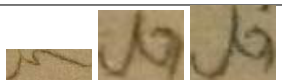
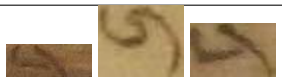
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Gantungan Ha			A	No
2	Gantungan Na			NA	No
3	Gantungan Ca			CA	No
4	Guwung / Cakra	Panganggе Aksara	Panganggе Aksara	RA	No
5	Gantungan Ka			KA	No
6	Gantungan Da			DA	No
7	Gantungan Ta			TA	No
8	Gempelan Sa		N/A	SA	No
9	Suku Kembung	Panganggе Aksara	Panganggе Aksara	WA	No
10	Gantungan La			LA	No
11	Gantungan Ma			MA	No
12	Gantungan Ga			GA	No
13	Gantungan Ba			BA	No
14	Gantungan Nga			NGA	No
15	Gempelan Pa			PA	No
16	Gantungan Ja			JA	No
17	Nania	Panganggе Aksara	Panganggе Aksara	YA	No
18	Gantungan Nya			NYA	No

Table C.3: ”Special” Consonants in Balinese Script

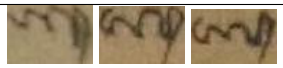
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Ka Mahaprana	ᬑ	N/A	KHA	Yes
2	Ga Gora	ᬖ	N/A	GHA	Yes
3	Ca Laca	ᬙ	N/A	CHA	Yes
4	Ja Jera	ᬛ	N/A	JHA	Yes
5	Sa Saga	ᬰ		SHA	Yes
6	Sa Sapa	ᬱ		SSA	Yes
7	Ta Latik	ᬟ	N/A	TTA	Yes
8	Ta Latik Mahaprana / Ta Murda Mahaprana	ᬠ	N/A	TTHA	Yes
9	Ta Tawa	ᬣ		THA	Yes
10	Da Madu	ᬥ		DHA	Yes
11	Da Murda Alpaprana	ᬡ	N/A	DDA	Yes
12	Da Murda Mahaprana	ᬢ	N/A	DDHA	Yes
13	Na Rambat	ᬦ᭄ᬭ		NNA	Yes
14	Pa Kapal	ᬯ	N/A	PHA	Yes
15	Ba Kembang	ᬪ		BHA	Yes

Table C.4: Conjunct Forms of ”Special” Consonants in Balinese Script

No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Gantungan Ka Mahaprana		N/A	KHA	No
2	Gantungan Ga Gora		N/A	GHA	No
3	Gantungan Ca Laca		N/A	CHA	No
4	Gantungan Ja Jera		N/A	JHA	No
5	Gantungan Sa Saga		N/A	SHA	No
6	Gempelan Sa Sapa			SSA	No
7	Gantungan Ta Latik			TTA	No
8	Gantungan Ta Latik Mahaprana / Gantungan Ta Murda Mahaprana		N/A	TTHA	No
9	Gantungan Ta Tawa			THA	No
10	Gantungan Da Madu		N/A	DHA	No
11	Gantungan Da Murda Alpaprana		N/A	DDA	No
12	Gantungan Da Murda Mahaprana		N/A	DDHA	No
13	Gantungan Na Rambat		N/A	NNA	No
14	Gempelan Pa Kapal		N/A	PHA	No
15	Gantungan Ba Kembang		N/A	BHA	No

Table C.5: "Basic" Independent Vowels in Balinese Script

No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	A Kara			A	Yes
2	I Kara			I	Yes
3	U Kara			U	Yes
4	E Kara			E	Yes
5	O Kara		N/A	O	Yes
6	Ra Repa		N/A	RE	Yes
7	La Lenga			LE	Yes

Table C.6: ”Longer” Independent Vowels in Balinese Script

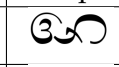
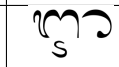
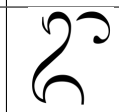
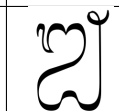
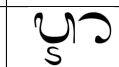
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	A Kara Tedung		N/A	A	Yes
2	I Kara Tedung		N/A	I	Yes
3	U Kara Tedung		N/A	U	Yes
4	Ai Kara / Airsanya		N/A	AI	Yes
5	O Kara Tedung / Au Kara		N/A	O	Yes
6	Ra Repa Tedung		N/A	RE	Yes
7	La Lenga Tedung		N/A	LE	Yes

Table C.7: Pangangge Suara (Dependent Vowels) in Balinese Script

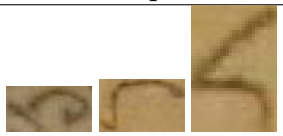
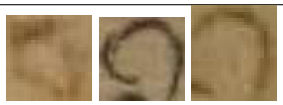
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Pepet				Yes
2	Ulu				Yes
3	Guwung Macelek				Yes
4	Gantungan La Pepet		N/A		Yes
5	Suku				Yes
6	Taling				Yes
7	Taling Tedung		N/A		Yes

Table C.8: ”Longer” Pangangge Suara (Dependent Vowels) in Balinese Script

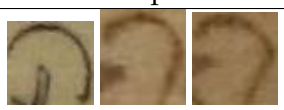
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Tedung	ᬾ			Yes
2	Ulu Sari	ᬿ			Yes
3	Guwung Macelek Tedung	ᬀ	N/A		Yes
4	Gantungan La Pepet Tedung	ᬁ	N/A		Yes
5	Suku Ilut	ᬅ			Yes
6	Taling Repa	ᬆ	N/A		Yes
7	Taling Repa Tedung	ᬇ	N/A		Yes

Table C.9: Pangangge Tengenan in Balinese Script

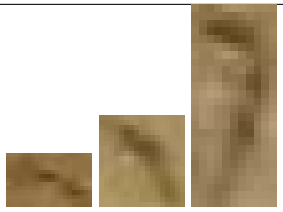
No	Name	Printed Sample	AMADI_Lontar Sample	Transliteration	UNICODE
1	Bisah				Yes
2	Surang				Yes
3	Cecek				Yes
4	Adeg-adeg				Yes

Table C.10: Pangangge Aksara in Balinese Script

No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Guwung / Cakra			RA	No
2	Suku Kembung			WA/UA	No
3	Nania			YA/IA	No

Table C.11: Digits in Balinese Script

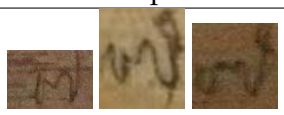
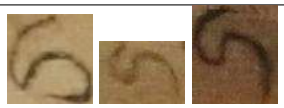
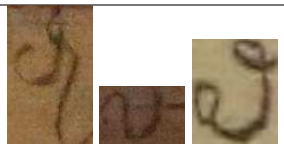
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	1	᭑		1	Yes
2	2	᭒		2	Yes
3	3	᭓		3	Yes
4	4	᭔		4	Yes
5	5	᭕		5	Yes
6	6	᭖		6	Yes
7	7	᭗		7	Yes
8	8	᭘		8	Yes
9	9	᭙		9	Yes
10	0	᭐		0	Yes

Table C.12: Additional Signs and Symbols in Balinese Script

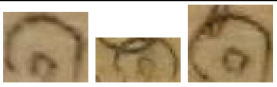
No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Ulu Ricem				Yes
2	Ulu Candra				Yes
3	Pepet Tedung		N/A		Yes
4	Pasalinan		N/A		No
5	Ongkara		N/A	OM	No

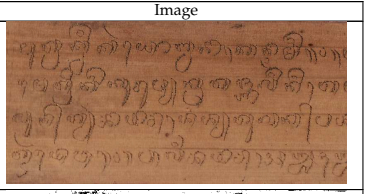
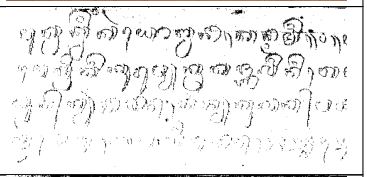
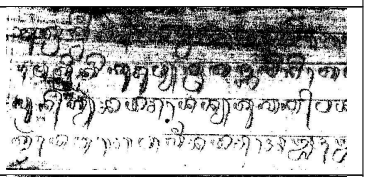
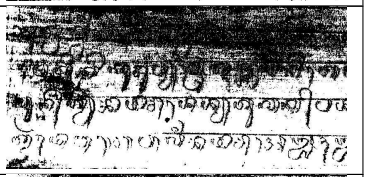
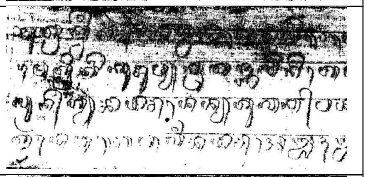
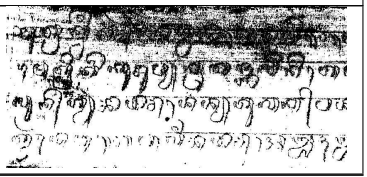
Table C.13: Punctuations in Balinese Script

No	Name	Printed Sample	AMADI_LontarSet Sample	Transliteration	UNICODE
1	Pamada				Yes
2	Panti		N/A		Yes
3	Carik Pamungkah		N/A	:	Yes
4	Carik Siki		Pangangge Tengenan Cecek	,	Yes
5	Carik Pareren		N/A	.	Yes
6	Windu		N/A		Yes
7	Pemeneng		N/A		Yes
8	Rerekan		N/A		Yes

Appendix D

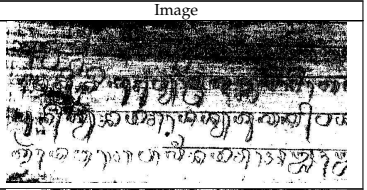
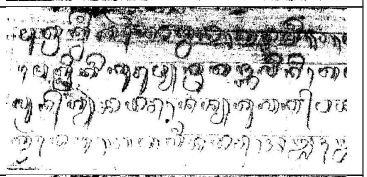
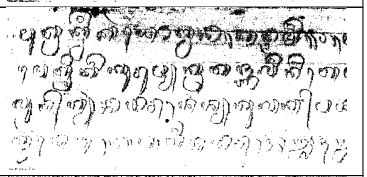
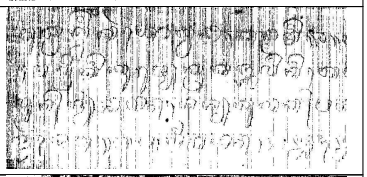
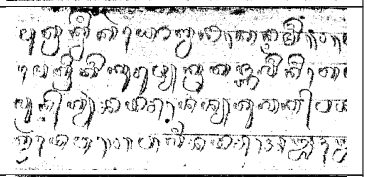
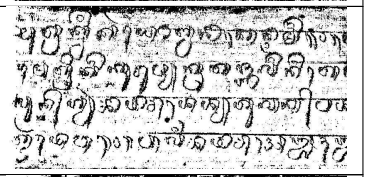
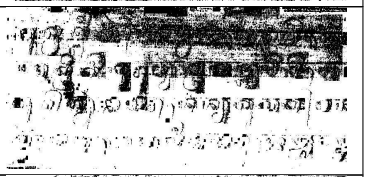
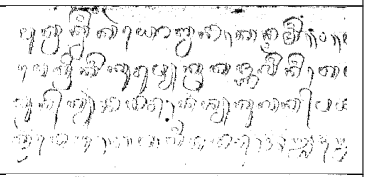
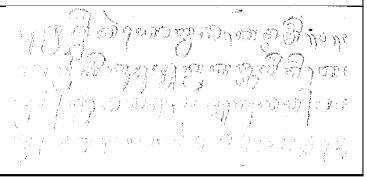
Results of Experiment : Comparison of Binarization Methods

Table D.1: Results of Experiment : Comparison of Binarization Methods for Document Images of Balinese Palm Leaf Manuscript

No	Method	Type	Parameter	Image
0	Original Image	-	-	
1.	Global Thresholding/ Global Fixed Threshold [6, 7, 8, 5]	Global	fixed threshold value = 122 (Empirically tested)	
2.	Global Thresholding/ Global Fixed Threshold [6, 7, 8, 5]	Global	fixed threshold value = 154 (Empirically tested)	
3.	Global Thresholding/ Global Fixed Threshold [6, 7, 8, 5]	Global	fixed threshold value = 159 (Mean of gray value of all pixels)	
4.	Global Means Iteration	Global	Mean initial = Mean of gray value of all pixels, threshold final founded = 152	
5.	K-Means Variation	Global	Mean initial 1 = mean of gray value of all pixels on image corner Mean initial 2 = mean of gray value of other pixels threshold final founded = 152	

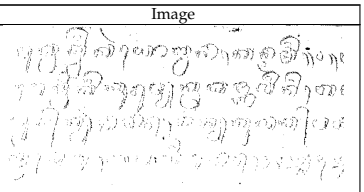
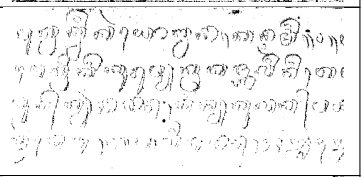
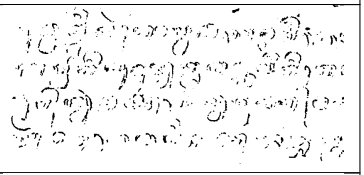
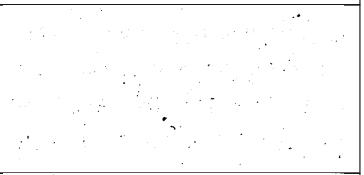
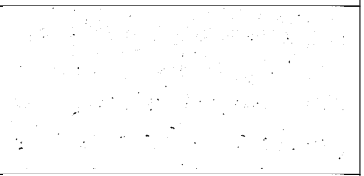
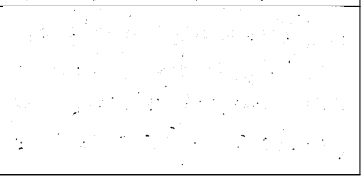
continued on next page

Table D.1 – continued from previous page

No	Method	Type	Parameter	Image
6.	Mixture Gaussian Model	Global	Initial threshold = Threshold Otsu threshold final founded = 159	
7.	Otsu [6, 5]	Global	No parameter threshold final founded = 146	
8.	Multiresolution Otsu (MROtsu) [6]	Global	Smallest block size = 200 (2xh) Dominant line height (h) = 100 (visually observed from image test) a hypothesized white/black ratio = 2:1 threshold final founded = 141	
9.	Chang [6]	Local	threshold_global = Threshold Otsu term "far away" from threshold_global = 0.25σ, where σ is the standard deviation of the full image window size = optimal scale for each pixel (calculated automatically)	
10.	Niblack [6, 7, 8, 4, 5]	Local	window size = 50 (visually observed from image test, big enough to cover a character) k = -0.2 (by the authors of Niblack)	
11.	Niblack [6, 7, 8, 4, 5]	Local	window size = 50 (visually observed from image test, big enough to cover a character) k = 0.2 (Empirically tested)	
12.	Niblack [6, 7, 8, 4, 5]	Local	window size = 50 (visually observed from image test, big enough to cover a character) k = 0 (Empirically tested)	
13.	Adaptive Niblack [7]	Local	threshold global = Threshold Otsu	
14.	Sauvola [6, 7, 8, 9, 4, 5]	Local	window_size = 50 (visually observed from image test, big enough to cover a character) k = 0.2 (Empirically tested) R = 128	
15.	Sauvola [6, 7, 8, 9, 4, 5]	Local	window_size = 50 (visually observed from image test, big enough to cover a character) k = 0.5, R = 128	

continued on next page

Table D.1 – continued from previous page

No	Method	Type	Parameter	Image
16.	Wolf [8, 4]	Local	window_size = 50 (visually observed from image test, big enough to cover a character) k = 0.5	
17.	Rais [5]	Local	window_size = 50 (visually observed from image test, big enough to cover a character)	
18.	NICK [4]	Local	window_size = 50 (visually observed from image test, big enough to cover a character) k = -0.2 (by the authors of Niblack)	
19.	Howe [10]		Base algorithm with static parameters	
20.	Howe [10]		Algorithm 1 (tune c)	
21.	Howe [10]		Algorithm 2 (tune c and t_hi)	
22.	Howe [10]		Algorithm 3 (tune c and pick t_hi from two alternatives)	

Appendix E

Compound Glyphs in AMADI_LontarSet

Table E.1: Compound Glyphs in AMADI_LontarSet

No	Name	AMADI_LontarSet Sample	Transliteration
1	TU		TU
2	KU		KU
3	I		I
4	NI		NI
5	TI		TI
6	U		U
7	RU		RU

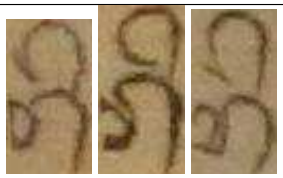
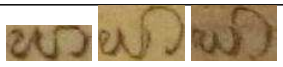
continued on next page

Table E.1 – continued from previous page

No	Name	AMADI_LontarSet Sample	Transliteration
8	DU		DU
9	WI		WI
10	DI		DI
11	KA-TEDONG		KA
12	WU		WU
13	GU		GU
14	MA-TEDONG		MA
15	NING		NING
16	KI		KI
17	NA-TEDONG		NA
18	LU		LU
19	BU		BU

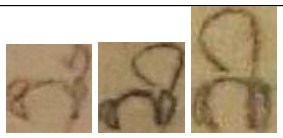
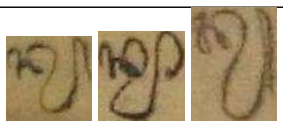
continued on next page

Table E.1 – continued from previous page

No	Name	AMADI_LontarSet Sample	Transliteration
20	YU		YU
21	TIA		TIA
22	JU		JU
23	NU		NU
24	RI		RI
25	TA-TEDONG		TA
26	LI		LI
27	MU		MU
28	YA-TEDONG		YA
29	SI		SI
30	DA-TEDONG		DA
31	PU		PU

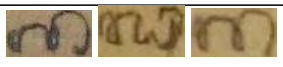
continued on next page

Table E.1 – continued from previous page

No	Name	AMADI_LontarSet Sample	Transliteration
32	RA-TEDONG		RA
33	KNA		KNA
34	SU		SU
35	GI		GI
36	NIA		NIA
37	NGU		NGU
38	KRA		KRA
39	JI		JI
40	JNA		JNA
41	WA-TEDONG		WA
42	BRA		BRA

continued on next page

Table E.1 – continued from previous page

No	Name	AMADI_LontarSet Sample	Transliteration
43	MI		MI
44	GNA		GNA
45	WRE		WRE
46	NA-RAMBAT-TEDONG		NA
47	JA-TEDONG		JA
48	BI		BI
49	PI		PI
50	TRA		TRA
51	NGI		NGI
52	LA-TEDONG		LA
53	WUA		WUA
54	GA-TEDONG		GA
55	CU		CU
56	A-TEDONG		A

continued on next page

Table E.1 – continued from previous page

No	Name	AMADI_LontarSet Sample	Transliteration
57	IA		IA
58	TUA		TUA
59	GRA		GRA

Appendix F

Glyph Dictionary and Example of XML file

Table F.1: Glyph Dictionary

No	Level1	Level2	Level3	Id	Sound	End	Split
1	TALENG	VOC	BASE-DESC	1	E	*	*
2	NA	CON	BASE	2	N	A	*
3	KA	CON	BASE	3	K	A	*
4	TA	CON	BASE	4	T	A	*
5	A	CON	BASE	5		A	*
6	ULU	VOC	ASC	6	I	*	*
7	CECEK	VOC	ASC	7	NG	*	*
8	WA	CON	BASE	8	W	A	*
9	DA	CON	BASE	9	D	A	*
10	ADEG-ADEG	GEM	ASC-BASE-DESC	10	*	*	*
11	JA	CON	BASE	11	J	A	*
12	BISAH	VOC	BASE-DESC	12	H	*	*
13	LA	CON	BASE	13	L	A	*
14	YA	CON	BASE	14	Y	A	*
15	MA	CON	BASE	15	M	A	*
16	SA	CON	BASE	16	S	A	*
17	TEDONG	VOC	BASE	17	*	*	*
18	SUKU	VOC	DESC	18	U	*	*
19	BA	CON	BASE	19	B	A	*
20	RA	CON	BASE	20	R	A	*
21	NGA	CON	BASE	21	NG	A	*
22	PEPET	VOC	ASC	22	E	*	*
23	SURANG	VOC	ASC	23	R	*	*
24	NANIA	VOC	BASE-DESC	24	I	A	*
25	GA	CON	BASE	25	G	A	*
26	TU	CON	BASE-DESC	26	TU	*	T_u
27	PA	CON	BASE	27	P	A	*
28	KU	CON	BASE-DESC	28	KU	*	K_u
29	GANTUNGAN TA	GAN	DESC	29	T	A	*
30	GANTUNGAN DA	GAN	DESC	30	D	A	*

continued on next page

Table F.1 – continued from previous page

No	Level1	Level2	Level3	Id	Sound	End	Split
31	GANTUNGAN A	GAN	DESC	31		A	*
32	GANTUNGAN MA	GAN	DESC	32	M	A	*
33	I	CON	ASC-BASE	33	I	*	A_i
34	GANTUNGAN KA	GAN	DESC	34	K	A	*
35	NI	CON	ASC-BASE	35	NI	*	N_i
36	DA MADU	CON	BASE	36	DH	A	*
37	TI	CON	ASC-BASE	37	TI	*	T_i
38	CA	CON	BASE	38	C	A	*
39	NA RAMBAT	CON	BASE	39	N	A	*
40	NYA	CON	BASE	40	NY	A	*
41	SUKU KEMBUNG	GAN	DESC	41	W	A	*
42	U KARA	CON	BASE-DESC	42	U	*	*
43	U	CON	BASE-DESC	43	U	*	A_u
44	RU	CON	BASE-DESC	44	RU	*	R_u
45	DU	CON	BASE-DESC	45	DU	*	D_u
46	SA SAGA	CON	BASE	46	S	A	*
47	WI	CON	ASC-BASE	47	WI	*	W_i
48	DI	CON	ASC-BASE	48	DI	*	D_i
49	KA TEDONG	CON	BASE	49	K	A	*
50	WU	CON	BASE-DESC	50	WU	*	W_u
51	GANTUNGAN LA	GAN	DESC	51	L	A	*
52	GUWUNG	GAN	DESC	52	R	A	*
53	GU	CON	BASE-DESC	53	GU	*	G_u
54	GANTUNGAN GA	GAN	DESC	54	G	A	*
55	MA TEDONG	CON	BASE	55	M	A	*
56	NING	CON	ASC-BASE	56	NING	*	N_ing
57	KI	CON	ASC-BASE	57	KI	*	K_i
58	SA SAPA	CON	BASE	58	S	A	*
59	GANTUNGAN NA	GAN	DESC	59	N	A	*
60	1	NUM	BASE	60	1	*	*
61	NA TEDONG	CON	BASE	61	N	A	*
62	SUKU ILUT	VOC	DESC	62	U	*	*
63	LU	CON	BASE-DESC	63	LU	*	L_u
64	0	NUM	BASE	64	0	*	*
65	BU	CON	BASE-DESC	65	BU	*	B_u
66	YU	CON	BASE-DESC	66	YU	*	Y_u
67	TIA	CON	BASE-DESC	67	TI	A	T_ia
68	JU	CON	BASE-DESC	68	JU	*	J_u
69	NU	CON	BASE-DESC	69	NU	*	N_u
70	BA KEMBANG	CON	BASE	70	BH	A	*
71	RI	CON	ASC-BASE	71	RI	*	R_i
72	TA TEDONG	CON	BASE	72	T	A	*
73	LI	CON	ASC-BASE	73	LI	*	L_i
74	MU	CON	BASE-DESC	74	MU	*	M_u
75	YA TEDONG	CON	BASE	75	Y	A	*
76	SI	CON	ASC-BASE	76	SI	*	S_i
77	GANTUNGAN TA LATIK	GAN	DESC	77	T	A	*

continued on next page

Table F.1 – continued from previous page

No	Level1	Level2	Level3	Id	Sound	End	Split
78	A KARA	CON	BASE	78		A	*
79	3	NUM	BASE-DESC	79	3	*	*
80	GEMPELAN SA SAPA	GEM	BASE	80	S	A	*
81	I KARA	CON	BASE-DESC	81	I	*	*
82	GANTUNGAN BA	GAN	DESC	82	B	A	*
83	DA TEDONG	CON	BASE	83	D	A	*
84	PU	CON	BASE-DESC	84	PU	*	P_u
85	4	NUM	BASE	85	4	*	*
86	RA TEDONG	CON	BASE	86	R	A	*
87	KNA	CON	BASE-DESC	87	KN	A	K_na
88	SU	CON	BASE-DESC	88	SU	*	S_u
89	GI	CON	ASC-BASE	89	GI	*	G_i
90	NIA	CON	BASE-DESC	90	NI	A	N_ia
91	2	NUM	BASE-DESC	91	2	*	*
92	NGU	CON	BASE-DESC	92	NGU	*	NG_u
93	KRA	CON	BASE-DESC	93	KR	A	K_ra
94	ULU CANDRA	VOC	ASC	94	I	*	*
95	JI	CON	ASC-BASE	95	JI	*	J_i
96	ULU SARI	VOC	ASC	96	I	*	*
97	JNA	CON	BASE-DESC	97	JN	A	J_na
98	WA TEDONG	CON	BASE	98	W	A	*
99	6	NUM	BASE	99	6	*	*
100	GANTUNGAN CA	GAN	DESC	100	C	A	*
101	GEMPELAN PA	GEM	BASE	101	P	A	*
102	9	NUM	BASE	102	9	*	*
103	BRA	CON	BASE-DESC	103	BR	A	B_ra
104	MI	CON	ASC-BASE	104	MI	*	M_i
105	5	NUM	BASE-DESC	105	5	*	*
106	GANTUNGAN JA	GAN	DESC	106	J	A	*
107	7	NUM	BASE	107	7	*	*
108	GNA	CON	BASE-DESC	108	GN	A	G_na
109	WRE	CON	BASE-DESC	109	WRE	*	W_re
110	GANTUNGAN NYA	GAN	DESC	110	NY	A	*
111	NA RAMBAT TEDONG	CON	BASE	111	N	A	*
112	LA LENGA	CON	BASE-DESC	112	LE	*	*
113	GUWUNG MACELEK	VOC	DESC	113	RE	*	*
114	JA TEDONG	CON	BASE	114	J	A	*
115	BI	CON	ASC-BASE	115	BI	*	B_i
116	PI	CON	ASC-BASE	116	PI	*	P_i
117	ULU RICEM	VOC	ASC	117	AM	*	*
118	GANTUNGAN NGA	GAN	DESC	118	NG	A	*
119	TRA	CON	BASE-DESC	119	TR	A	T_ra
120	E KARA	CON	BASE	120	E	*	*
121	NGI	CON	ASC-BASE	121	NGI	*	NG_i
122	LA TEDONG	CON	BASE	122	L	A	*
123	WUA	CON	BASE-DESC	123	WU	A	W_ua
124	GA TEDONG	CON	BASE	124	G	A	*

continued on next page

Table F.1 – continued from previous page

No	Level1	Level2	Level3	Id	Sound	End	Split
125	8	NUM	BASE	125	8	*	*
126	PAMADA	PUN	BASE-DESC	126	*	*	*
127	CU	CON	BASE-DESC	127	CU	*	C_u
128	A TEDONG	CON	BASE	128		A	*
129	GANTUNGAN TA TAWA	GAN	DESC	129	T	A	*
130	IA	CON	BASE-DESC	130	I	A	A_ia
131	TUA	CON	BASE-DESC	131	TU	A	T_ua
132	TA TAWA	CON	BASE	132	T	A	*
133	GRA	CON	BASE-DESC	133	GR	A	G_ra

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<BIBLIOGRAPHY>
<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>TALENG</NAME>
        <LEVEL>Level2</LEVEL> <NAME>VOC</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE-DESC</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>1</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>E</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>*</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>NA</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>2</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>N</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>KA</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>3</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>K</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>TA</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>4</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>T</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>A</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
```

```
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>5</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE> </VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>ULU</NAME>
        <LEVEL>Level2</LEVEL> <NAME>VOC</NAME>
        <LEVEL>Level3</LEVEL> <NAME>ASC</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>6</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>I</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>*</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>CECEK</NAME>
        <LEVEL>Level2</LEVEL> <NAME>VOC</NAME>
        <LEVEL>Level3</LEVEL> <NAME>ASC</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>7</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>NG</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>*</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>WA</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>8</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>W</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>

<OBJECT>
    <CLASS>
        <LEVEL>Level1</LEVEL> <NAME>DA</NAME>
        <LEVEL>Level2</LEVEL> <NAME>CON</NAME>
        <LEVEL>Level3</LEVEL> <NAME>BASE</NAME>
    </CLASS>
    <PARAMETER>id</PARAMETER> <VALUE>9</VALUE>
    <PARAMETER>sound</PARAMETER> <VALUE>D</VALUE>
    <PARAMETER>end</PARAMETER> <VALUE>A</VALUE>
    <PARAMETER>split</PARAMETER> <VALUE>*</VALUE>
</OBJECT>
```

Appendix G

Phonological Rules for Transliteration of Balinese Script

- **RULE1**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE ⇒ SPEECH_SOUND = CURR.BASE.STARTSYLLABLE
- **RULE2**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = SPEECH_SOUND + CURR.DESC.STARTSYLLABLE
- **RULE3**: IF PREV.BASE.LEVEL1 = ″TALENG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL2 ≠ GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ″TEDONG″ ⇒ SPEECH_SOUND = SPEECH_SOUND + PREV.BASE.STARTSYLLABLE
- **RULE4**: IF PREV.BASE.LEVEL1 = ″TALENG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL2 ≠ GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 = ″TEDONG″ ⇒ SPEECH_SOUND = SPEECH_SOUND + ″O″
- **RULE5**: IF PREV2.BASE.LEVEL1 = ″TALENG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 ≠ CON AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ″TEDONG″ ⇒ SPEECH_SOUND = SPEECH_SOUND + ″E″
- **RULE6**: IF PREV2.BASE.LEVEL1 = ″TALENG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 ≠ CON AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 = ″TEDONG″ ⇒ SPEECH_SOUND = SPEECH_SOUND + ″O″
- **RULE7**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 ≠ ″NANIA″ ⇒ SPEECH_SOUND = SPEECH_SOUND + NEXT.BASE.STARTSYLLABLE
- **RULE8**: IF PREV.BASE.LEVEL1 ≠ ″TALENG″ AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = BASE AND CURR.BASE.ENDSYLLABLE = ″A″ AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ″ADEG-ADEG″ AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ″A″
- **RULE9**: IF PREV.BASE.LEVEL1 ≠ ″TALENG″ AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.BASE.ENDSYLLABLE = ″A″ AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ″ADEG-ADEG″ AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ″A″
- **RULE10**: IF PREV.BASE.LEVEL1 ≠ ″TALENG″ AND CURR.ASC.LEVEL1 = ″CECEK″ / ″SURANG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = BASE AND CURR.BASE.ENDSYLLABLE = ″A″ AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ″ADEG-ADEG″ AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ″A″
- **RULE11**: IF PREV.BASE.LEVEL1 ≠ ″TALENG″ AND CURR.ASC.LEVEL1 = ″CECEK″ / ″SURANG″ AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.BASE.ENDSYLLABLE = ″A″ AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ″ADEG-ADEG″ AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ″A″
- **RULE12**: IF PREV.BASE.LEVEL1 ≠ ″TALENG″ AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON/GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.DESC.LEVEL1 ≠ EMPTY AND CURR.DESC.STARTSYLLABLE = ″A″ AND NEXT.BASE.LEVEL1 ≠ ″ADEG-ADEG″ AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ″A″

- **RULE13**: IF PREV.BASE.LEVEL1 ≠ ”TALENG” AND CURR.ASC.LEVEL1 = ”CECEK” / ”SURANG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON/GEM AND CURR.BASE.LEVEL3 = BASE AND CURR.DESC.LEVEL1 ≠ EMPTY AND CURR.DESC.STARTSYLLABLE = ”A” AND NEXT.BASE.LEVEL1 ≠ ”ADEG-ADEG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”A”

- **RULE14**: IF CURR.ASC.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE ⇒ SPEECH_SOUND = SPEECH_SOUND + CURR.ASC.STARTSYLLABLE

- **RULE15**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON / GEM AND CURR.BASE.LEVEL3 = BASE AND NEXT.BASE.LEVEL1 = ”BISAH” ⇒ SPEECH_SOUND = SPEECH_SOUND + NEXT.BASE.STARTSYLLABLE

- **RULE16a**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND (CURR.BASE.SPLITSYLLABLE ≠ ”*” ⇒ ROOT_PART = SPLIT_PART1(CURR.BASE.SPLITSYLLABLE)) AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = ROOT_PART

- **RULE16b**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND (CURR.BASE.SPLITSYLLABLE = ”*” ⇒ ROOT_PART = ””) AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = ROOT_PART

- **RULE17**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND CURR.DESC.LEVEL1 = EMPTY ⇒ SPEECH_SOUND = CURR.BASE.STARTSYLLABLE

- **RULE18**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = SPEECH_SOUND + CURR.DESC.STARTSYLLABLE

- **RULE19**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 ≠ ”TEDONG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + PREV.BASE.STARTSYLLABLE

- **RULE20**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 = ”TEDONG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”O”

- **RULE21**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 ≠ ”TEDONG” AND NEXT.BASE.LEVEL2 = GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”E”

- **RULE22**: IF PREV.BASE.LEVEL1 = ”TALENG” AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 = ”TEDONG” AND NEXT.BASE.LEVEL2 = GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”O”

- **RULE23**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 = ”NANIA” ⇒ SPEECH_SOUND = SPEECH_SOUND + NEXT.BASE.STARTSYLLABLE

- **RULE24**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND (CURR.BASE.SPLITSYLLABLE ≠ ”*” ⇒ VOC_PART = SPLIT_PART2(CURR.BASE.SPLITSYLLABLE)) AND CURR.DESC.LEVEL1 ≠ EMPTY ⇒ SPEECH_SOUND = SPEECH_SOUND + VOC_PART

- **RULE25**: IF PREV.BASE.LEVEL1 ≠ ”TALENG” AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND (CURR.BASE.SPLITSYLLABLE = ”*” ⇒ VOC_PART = ””) AND CURR.DESC.LEVEL1 ≠ EMPTY AND CURR.DESC.STARTSYLLABLE = ”A” AND NEXT.BASE.LEVEL1 ≠ ”ADEG-ADEG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”A”

- **RULE26**: IF PREV.BASE.LEVEL1 ≠ ”TALENG” AND CURR.ASC.LEVEL1 = EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND CURR.BASE.ENDSYLLABLE = ”A” AND CURR.DESC.LEVEL1 = EMPTY AND NEXT.BASE.LEVEL1 ≠ ”ADEG-ADEG” AND NEXT.BASE.LEVEL2 ≠ GEM ⇒ SPEECH_SOUND = SPEECH_SOUND + ”A”

- **RULE27**: IF CURR.ASC.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC ⇒ SPEECH_SOUND = SPEECH_SOUND + CURR.ASC.STARTSYLLABLE

- **RULE28**: IF CURR.BASE.LEVEL1 ≠ EMPTY AND CURR.BASE.LEVEL2 = CON AND CURR.BASE.LEVEL3 = ASC-BASE / BASE-DESC AND NEXT.BASE.LEVEL1 = ”BISAH” ⇒ SPEECH_SOUND = SPEECH_SOUND + NEXT.BASE.STARTSYLLABLE

- **RULE29**: IF PREV.BASE.LEVEL1 = ”TEDONG” AND CURR.BASE.LEVEL1 = ”BISAH” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”H”

- **RULE30**: IF CURR.BASE.LEVEL1 = EMPTY ⇒ SPEECH_SOUND = SPEECH_SOUND + ”*”

- **RULE31**: IF CURR.BASE.LEVEL1 = ”CECEK” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”,”
- **RULE32**: IF CURR.BASE.LEVEL1 = ”TALENG” / ”TEDONG” / ”NANIA” / ”BISAH” / ”ADEG-ADEG” ⇒ SPEECH_SOUND = SPEECH_SOUND + ”*”
- **RULE33**: IF CURR.BASE.LEVEL2 = NUM ⇒ SPEECH_SOUND = CURR.BASE.STARTSYLLABLE
- **RULE34**: IF CURR.BASE.LEVEL2 = PUN ⇒ SPEECH_SOUND = CURR.BASE.STARTSYLLABLE

The following Table G.1 summarizes all phonological rules in a table view format.

Table G.1 Phonological rules for the transliteration of the Balinese script

NO	PREV2	PREV	CURR								NEXT		RESULT
	BASE	BASE	ASC	BASE					DESC		BASE		
	LEVEL1	LEVEL1	LEVEL1	LEVEL1	LEVEL2	LEVEL3	END	SPLIT	LEVEL1	SOUND	LEVEL1	LEVEL2	
1				~EMPTY	CON/GEM	BASE							result=CURR_SOUND_BASE
2				~EMPTY	CON/GEM	BASE			~EMPTY				result=result+CURR_SOUND_DESC
3		TALENG		~EMPTY	CON&~GEM	BASE					~TEDONG		result=result+PREV_SOUND_BASE
4		TALENG		~EMPTY	CON&~GEM	BASE					TEDONG		result=result+"o"
5	TALENG			~EMPTY	~CON&GEM	BASE					~TEDONG		result=result+"e"
6	TALENG			~EMPTY	~CON&GEM	BASE					TEDONG		result=result+"o"
7				~EMPTY	CON/GEM	BASE					NANIA		result=result+NEXT_SOUND_BASE
8		~TALENG	EMPTY	~EMPTY	CON	BASE	"A"		EMPTY		~ADEG-ADEG	~GEM	result=result+"a"
9		~TALENG	EMPTY	~EMPTY	GEM	BASE	"A"		EMPTY		~ADEG-ADEG	~GEM	result=result+"a"
10		~TALENG	CECEK/SURANG	~EMPTY	CON	BASE	"A"		EMPTY		~ADEG-ADEG	~GEM	result=result+"a"
11		~TALENG	CECEK/SURANG	~EMPTY	GEM	BASE	"A"		EMPTY		~ADEG-ADEG	~GEM	result=result+"a"
12		~TALENG	EMPTY	~EMPTY	CON/GEM	BASE			~EMPTY	"A"	~ADEG-ADEG	~GEM	result=result+"a"
13		~TALENG	CECEK/SURANG	~EMPTY	CON/GEM	BASE			~EMPTY	"A"	~ADEG-ADEG	~GEM	result=result+"a"
14			~EMPTY	~EMPTY	CON/GEM	BASE							result=result+CURR_SOUND_ASC
15				~EMPTY	CON/GEM	BASE					BISAH		result=result+NEXT_SOUND_BASE
16				~EMPTY	CON	ASC-BASE/BASE-DESC		~"*", root_part=split_part1(CURR_SPLIT_BASE)	~EMPTY				result=root_part
				~EMPTY	CON	ASC-BASE/BASE-DESC		"*", root_part=""	~EMPTY				result=root_part
17				~EMPTY	CON	ASC-BASE/BASE-DESC			EMPTY				result=CURR_SOUND_BASE
18				~EMPTY	CON	ASC-BASE/BASE-DESC			~EMPTY				result=result+CURR_SOUND_DESC
19		TALENG		~EMPTY	CON	ASC-BASE/BASE-DESC					~TEDONG	~GEM	result=result+PREV_SOUND_BASE
20		TALENG		~EMPTY	CON	ASC-BASE/BASE-DESC					TEDONG	~GEM	result=result+"o"
21		TALENG		~EMPTY	CON	ASC-BASE/BASE-DESC					~TEDONG	GEM	result=result+"e"
22		TALENG		~EMPTY	CON	ASC-BASE/BASE-DESC					TEDONG	GEM	result=result+"o"
23				~EMPTY	CON	ASC-BASE/BASE-DESC					NANIA		result=result+NEXT_SOUND_BASE
24				~EMPTY	CON	ASC-BASE/BASE-DESC		~"*", voc_part=split_part2(CURR_SPLIT_BASE)	~EMPTY				result=result+voc_part
25		~TALENG	EMPTY	~EMPTY	CON	ASC-BASE/BASE-DESC		"*", voc_part=""	~EMPTY	"A"	~ADEG-ADEG	~GEM	result=result+"a"
26		~TALENG	EMPTY	~EMPTY	CON	ASC-BASE/BASE-DESC	"A"		EMPTY		~ADEG-ADEG	~GEM	result=result+"a"
27			~EMPTY	~EMPTY	CON	ASC-BASE/BASE-DESC							result=result+CURR_SOUND_ASC

28				~EMPTY	CON	ASC-BASE/BASE-DESC					BISAH		result=result+NEXT_SOUND_BASE
29		TEDONG		BISAH									result=result+"h"
30				EMPTY									result=result+"*"
31				CECEK									result=result+","
32				TALENG/TEDONG/NANIA/BISAH/ADEG-ADEG									result=result+"*"
33					NUM								result=CURR_SOUND_BASE
34					PUN								result=CURR_SOUND_BASE

Bibliography

[1] Made Windu Antara Kesiman, Sophea Prum, Jean-Christophe Burie, and Jean-Marc Ogier. An Initial Study On The Construction Of Ground Truth Binarized Images Of Ancient Palm Leaf Manuscripts. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, Nancy, France, August 2015.

[2] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. ICDAR 2013 Document Image Binarization Contest (DIBCO 2013). In *ICDAR '13 Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, pages 1471–1476. IEEE, August 2013.

[3] Ines Ben Messaoud, Haikal El Abed, Volker Märgner, and Hamid Amiri. A design of a preprocessing framework for large database of historical documents. In *HIP '11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, page 177. ACM Press, 2011.

[4] Khurram Khurshid, Imran Siddiqi, Claudie Faure, and Nicole Vincent. Comparison of Niblack inspired binarization methods for ancient documents. In *Proc. SPIE 7247, Document Recognition and Retrieval XVI*, volume 7247, pages 72470U–72470U–9, 2009.

[5] M. Shebzad Hanif Naveed Bin Rais. Adaptive thresholding technique for document image analysis. In *8th International Multitopic Conference, 2004. Proceedings of INMIC 2004.*, pages 61–66. IEEE, 2004.

[6] Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia. OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40(2):389–397, February 2007.

[7] J. He, Q.D.M. Do, A.C. Downton, and J.H. Kim. A comparison of binarization methods for historical archive documents. In *Eighth International Conference on Document Analysis and Recognition, 2005. Proceedings*, pages 538–542 Vol. 1, 2005.

[8] Meng-Ling Feng and Yap-Peng Tan. Contrast adaptive binarization of low quality document images. *IEICE Electronics Express*, 1(16):501–506, 2004.

[9] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[10] Nicholas R. Howe. Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(3):247–258, September 2013.

[11] Made Windu Antara Kesiman, Jean-Christophe Burie, and Jean-Marc Ogier. A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, pages 325–330, Shenzhen, China.

[12] Konstantinos Ntirogiannis, B. Gatos, and I. Pratikakis. An Objective Evaluation Methodology for Document Image Binarization Techniques. In *The Eighth IAPR International Workshop on Document Analysis Systems, 2008*, pages 217–224. IEEE, September 2008.

[13] K. Ntirogiannis, B. Gatos, and I. Pratikakis. Performance Evaluation Methodology for Historical Document Image Binarization. *IEEE Transactions on Image Processing*, 22(2):595–609, February 2013.

[14] Made Windu Antara Kesiman, Sophea Prum, Jean-Christophe Burie, and Jean-Marc Ogier. Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images. In *23rd International Conference on Pattern Recognition*, Cancun, Mexico, December 2016.

[15] Made Windu Antara Kesiman, Jean-Christophe Burie, Jean-Marc Ogier, Gusti Ngurah Made Agus Wibawantara, and I Made Gede Sunarya. AMADI_lontarset: The First Handwritten Balinese Palm Leaf Manuscripts Dataset. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, pages 168–172, Shenzhen, China, October 2016.

[16] Made Windu Antara Kesiman, Jean-Christophe Burie, and Jean-Marc Ogier. A Complete Scheme Of Spatially Categorized Glyph Recognition For The Transliteration Of Balinese Palm Leaf Manuscripts. In *14th IAPR International Conference on Document Analysis and Recognition*, Kyoto, Japan, November 2017.

[17] Eric Saund, Jing Lin, and Prateek Sarkar. PixLabeler: User Interface for Pixel-Level Labeling of Elements in Document Images. In *ICDAR '09. 10th International Conference on Document Analysis and Recognition*, pages 646–650. IEEE, 2009.

[18] Made Windu Antara Kesiman, Sophea Prum, I Made Gede Sunarya, Jean-Christophe Burie, and Jean-Marc Ogier. An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts. In *5th International Conference on Image Processing, Theory, Tools and Applications (IPTA) 2015*, pages 229–233, Orleans, France, November 2015.

[19] Made Windu Antara Kesiman. Extraction des chaînes des symboles dans les séquences. Master's thesis, Universite de La Rochelle, La Rochelle, France, July 2006.

[20] David Doermann and Karl Tombre, editors. *Handbook of Document Image Processing and Recognition*. Springer London, London, 2014.

[21] Rapeeporn Chamchong, Chun Che Fung, and Kok Wai Wong. Comparing Binarisation Techniques for the Processing of Ancient Manuscripts. In Ryohei Nakatsu, Naoko Tosa, Fazel Naghdy, Kok Wai Wong, and Philippe Codognet, editors, *Cultural Computing*, volume 333, pages 55–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[22] Chun Che Fung and R. Chamchong. A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts. In *Third International Conference on Knowledge Discovery and Data Mining, WKDD 2010*, pages 236–240. IEEE, January 2010.

[23] Jean-Christophe Burie, Mickaël Coustaty, Setiawan Hadi, Made Windu Antara Kesiman, Jean-Marc Ogier, Erick Paulus, Kimheng Sok, I Made Gede Sunarya, and Dona Valy. ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, pages 596–601, Shenzhen, China, October 2016.

[24] Made Kesiman, Dona Valy, Jean-Christophe Burie, Erick Paulus, Mira Suryani, Setiawan Hadi, Michel Verleysen, Sophea Chhun, and Jean-Marc Ogier. Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia. *Journal of Imaging*, 4(2):43, February 2018.

[25] Made Windu Antara Kesiman, Dona Valy, Jean-Christophe Burie, Erick Paulus, I. Made Gede Sunarya, Setiawan Hadi, Kim Heng Sok, and Jean-Marc Ogier. Southeast Asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges. *Journal of Electronic Imaging*, 26(1):011011, November 2016.

[26] Made Windu Antara Kesiman, Jean Christophe Burie, Jean Marc Ogier, and Philippe Grangé. Knowledge Representation and Phonological Rules for the Automatic Transliteration of Balinese Script on Palm Leaf Manuscript. *Computación y Sistemas*, 21(4), January 2018.

[27] Dr Balamurugan, Sangeetha K, and Dr P Sengottuvelan. Document image analysis -a review. *International Journal of Computer Applications*, 1(1):21, October 2011.

[28] Andrei Tigora. An overview of document image analysis system. *Journal of Information Systems and Operations Management*, 7:378–390, 12 2013.

[29] Martin Jenckel, Syed Saqib Bukhari, and Andreas Dengel. anyOCR: A sequence learning based OCR system for unlabeled historical documents. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4035–4040. IEEE, December 2016.

[30] Lawrence O'Gorman and Rangachar Kasturi. *Executive briefing: document image analysis*. IEEE Computer Society Press, Los Alamitos, Calif, 1997.

[31] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, January 2000.

[32] Rangachar Kasturi, Lawrence O'Gorman, and Venu Govindaraju. Document image analysis: A primer. *Sadhana*, 27(1):3–22, February 2002.

[33] R. J. Ramteke. Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition. *International Journal of Computer Applications*, 1(18):1–5, February 2010.

[34] Kartar Singh Siddharth, Renu Dhir, and Rajneesh Rani. Handwritten Gurmukhi Numeral Recognition using Different Feature Sets. *International Journal of Computer Applications*, 28(2):20–24, August 2011.

[35] Dharamveer Sharma and Puneet Jhajj. Recognition of Isolated Handwritten Characters in Gurmukhi Script. *International Journal of Computer Applications*, 4(8):9–17, August 2010.

[36] Ashutosh Aggarwal, Karamjeet Singh, and Kamalpreet Singh. Use of Gradient Technique for Extracting Features from Handwritten Gurmukhi Characters and Numerals. *Procedia Computer Science*, 46:1716–1723, 2015.

[37] G.S. Lehal and C. Singh. A Gurmukhi script recognition system. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 557–560. IEEE Comput. Soc, 2000.

[38] L. Rothacker, G. A. Fink, P. Banerjee, U. Bhattacharya, and B. B. Chaudhuri. Bag-of-features HMMs for segmentation-free Bangla word spotting. In *MOCR '13 Proceedings of the 4th International Workshop on Multilingual OCR*, page 1. ACM Press, 2013.

[39] Ashlin Deepa R.N and R.Rajeswara Rao. Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, 3(1):481–485, 2014.

[40] Aditia Gunawan. Nipah or Gebang? A Philological and Codicological Study Based on Sources from West Java. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 171:249–280, 2015.

[41] The Unicode® Standard Version 9.0 – Core Specification. Technical report, The Unicode Consortium, July 2016.

[42] Mira Suryani, Erick Paulus, Setiawan Hadi, Undang A. Darsa, and Jean-Christophe Burie. The Handwritten Sundanese Palm Leaf Manuscript Dataset From 15th Century. In *14th IAPR International Conference on Document Analysis and Recognition*, Kyoto, Japan, November 2017.

[43] Andrea Acri. Palm-leaf manuscripts in today's Bali. *Inside Indonesia*, September 2015.

[44] Andrea Acri. Living Balinese heritage: Palm-leaf manuscripts and their caretakers. *The Newsletter*, page 43, 2013.

[45] Helen Creese. Old Javanese legal traditions in pre-colonial Bali. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 165(2-3):241–290, 2009.

[46] H.I.R. Hinzler. Balinese palm-leaf manuscripts. *Bijdragen tot de taal-, land- en volkenkunde / Journal of the Humanities and Social Sciences of Southeast Asia*, 149(3):438–473, January 1993.

[47] G. E. Marrison. Balinese manuscripts in the Library of the University of Leiden and other collections in the Netherlands: a review article. *Journal of the Royal Asiatic Society of Great Britain & Ireland*, 120(02):378–384, April 1988.

[48] Ida Bagus Adi Sudewa. Contemporary Use of The Balinese Script. Technical Report L2/03-118, https://docplayer.net/31124487-Contemporary-use-of-the-balinese-script.html, December 2003.

[49] Rapeeporn Chamchong and Chun Che Fung. Character segmentation from ancient palm leaf manuscripts in Thailand. In *HIP '11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, page 140. ACM Press, 2011.

[50] Rapeeporn Chamchong and Chun Che Fung. Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 588–593. IEEE, September 2012.

[51] Adnan Ul-Hasan, Syed Saqib Bukhari, and Andreas Dengel. OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 174–179. IEEE, April 2016.

[52] Anguelos Nicolaou and Basilis Gatos. Handwritten Text Line Segmentation by Shredding Text into its Lines. In *2009 10th International Conference on Document Analysis and Recognition*, pages 626–630. IEEE, 2009.

[53] N. Arica and F.T. Yarman-Vural. Optical character recognition for cursive handwriting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):801–813, June 2002.

[54] M. Blumenstein, B. Verma, and H. Basli. A novel feature extraction technique for the recognition of segmented handwritten characters. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, volume 1, pages 137–141. IEEE Comput. Soc, 2003.

[55] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J.Heard. Building a test collection for complex document information processing. In *Proc. 29th Annual Int. ACM SIGIR Conference*, pages 665–666, 2006.

[56] G. Agam, S. Argamon, O. Frieder, D. Grossman, and D. Lewis. *The Complex Document Image Processing (CDIP) Test Collection Project*. Illinois Institute of Technology, 2006.

[57] University of California, San Francisco. *The Legacy Tobacco Document Library (LTDL)*, 2007.

[58] Berrin A. Yanikoglu and Luc Vincent. PINK PANTHER: A Complete Environment for Ground-Truthing and Benchmarking Document Page Segmentation. *Pattern Recognition*, 31(9):1191–1204, September 1998.

[59] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments. In *2011 International Conference on Document Analysis and Recognition*, pages 48–52. IEEE, September 2011.

[60] Róisín Rowley-Brooke, François Pitié, and Anil Kokaram. A Ground Truth Bleed-Through Document Image Database. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides, editors, *Theory and Practice of Digital Libraries*, volume 7489, pages 185–196. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[61] Hossein Ziaei Nafchi, Seyed Morteza Ayatollahi, Reza Farrahi Moghaddam, and Mohamed Cheriet. An Efficient Ground Truthing Tool for Binarization of Historical Manuscripts. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 807–811. IEEE, August 2013.

[62] G. Bal, G. Agam, O. Frieder, and G. Frieder. Interactive degraded document enhancement and ground truth generation. In Berrin A. Yanikoglu and Kathrin Berkner, editors, *Proceedings Volume 6815, Document Recognition and Retrieval XV; 68150Z (2008)*, volume 6815, pages 68150Z–68150Z–9, January 2008.

[63] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, November 1986.

[64] Dona Valy, Michel Verleysen, and Kimheng Sok. Line Segmentation for Grayscale Text Images of Khmer Palm Leaf Manuscripts. In *7th International Conference on Image Processing Theory, Tools and Applications (IPTA 2017)*, Montreal, Canada, December 2017.

[65] C. Wolf, J.-M. Jolion, and F. Chassaing. Text localization, enhancement and binarization in multimedia documents. In *Object recognition supported by user interaction for service robots*, volume 2, pages 1037–1040. IEEE Comput. Soc, 2002.

[66] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Online and offline handwritten Chinese character recognition: Benchmarking on new databases. *Pattern Recognition*, 46(1):155–162, January 2013.

[67] Shijian Lu, Bolan Su, and Chew Lim Tan. Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(4):303–314, December 2010.

[68] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust Document Image Binarization Technique for Degraded Document Images. *IEEE Transactions on Image Processing*, 22(4):1408–1417, April 2013.

[69] N. Arica and F.T. Yarman-Vural. A new scheme for off-line handwritten connected digit recognition. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, volume 2, pages 1127–1129. IEEE Comput. Soc, 1998.

[70] Seong-Whan Lee, Dong-June Lee, and Hee-Seon Park. A new methodology for gray-scale character segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):1045–1050, October 1996.

[71] Zhixin Shi, S. Setlur, and V. Govindaraju. Text extraction from gray scale historical document images using adaptive local connectivity map. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 794–798 Vol. 2. IEEE, 2005.

[72] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis. Text line and word segmentation of handwritten documents. *Pattern Recognition*, 42(12):3169–3183, December 2009.

[73] Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang Ing Ren, and George D.C. Cavalcanti. Text Line Segmentation Based on Morphology and Histogram Projection. In *2009 10th International Conference on Document Analysis and Recognition*, pages 651–655. IEEE, 2009.

[74] Xi Zhang and Chew Lim Tan. Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 98–103. IEEE, September 2014.

[75] Fei Yin and Cheng-Lin Liu. Handwritten Chinese text line segmentation by clustering with distance metric learning. *Pattern Recognition*, 42(12):3146–3157, December 2009.

[76] Jayant Kumar, Wael Abd-Almageed, Le Kang, and David Doermann. Handwritten Arabic text line segmentation using affinity propagation. In *DAS '10 Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 135–142. ACM Press, 2010.

[77] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):123–138, April 2007.

[78] Itay Bar-Yosef, Nate Hagbi, Klara Kedem, and Itshak Dinstein. Line Segmentation for Degraded Handwritten Historical Documents. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1161–1165. IEEE, 2009.

[79] Angelika Garz, Andreas Fischer, Robert Sablatnig, and Horst Bunke. Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 95–99. IEEE, March 2012.

[80] Rapeeporn Chamchong and Chun Che Fung. Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 588–593. IEEE, September 2012.

[81] Olarik Surinta, Michiel Holtkamp, Faik Karabaa, Jean-Paul Van Oosten, Lambert Schomaker, and Marco Wiering. A Path Planning for Line Segmentation of Handwritten Documents. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 175–180. IEEE, September 2014.

[82] Nikolaos Arvanitopoulos and Sabine Susstrunk. Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 726–731. IEEE, September 2014.

[83] Raid Saabni and Jihad El-Sana. Language-Independent Text Lines Extraction Using Seam Carving. In *2011 International Conference on Document Analysis and Recognition*, pages 563–568. IEEE, September 2011.

[84] Abedelkadir Asi, Raid Saabni, and Jihad El-Sana. Text line segmentation for gray scale historical document images. In *HIP '11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, page 120. ACM Press, 2011.

[85] Christopher Stoll, Yingcai Xiao, and Zhong-Hui Duan. Text Line Extraction Using Seam Carving. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition 2015*.

[86] M. Zahid Hossain, M. Ashraful Amin, and Hong Yan. Rapid Feature Extraction for Optical Character Recognition. *CoRR*, abs/1206.0238, 2012.

[87] Satish Kumar. Neighborhood Pixels Weights-A New Feature Extractor. *International Journal of Computer Theory and Engineering*, pages 69–77, 2009.

[88] M. Bokser. Omnidocument technologies. *Proceedings of the IEEE*, 80(7):1066–1078, July 1992.

[89] Y. Fujisawa, Meng Shi, T. Wakabayashi, and F. Kimura. Handwritten numeral recognition using gradient and curvature of gray scale image. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, pages 277–280. IEEE, 1999.

[90] Zhen Jin, Kaiyue Qi, Yi Zhou, Kai Chen, Jianbo Chen, and Haibing Guan. SSIFT: An Improved SIFT Descriptor for Chinese Character Recognition in Complex Images. In *2009 International Symposium on Computer Network and Multimedia Technology*, pages 1–5. IEEE, December 2009.

[91] Manju Rani and Yogesh Kumar Meena. An Efficient Feature Extraction Method for Handwritten Character Recognition. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Bijaya Ketan Panigrahi, Ponnuthurai Nagaratnam Suganthan, Swagatam Das, and Suresh Chandra Satapathy, editors, *Swarm, Evolutionary, and Memetic Computing*, volume 7077, pages 302–309. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[92] Øivind Due Trier, Anil K. Jain, and Torfinn Taxt. Feature extraction methods for character recognition-A survey. *Pattern Recognition*, 29(4):641–662, April 1996.

[93] Satish Kumar. Study of Features for Hand-printed Recognition. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 5(12), 2011.

[94] Neha J. Pithadia and Dr. Vishal D. Nimavat. A Review on Feature Extraction Techniques for Optical Character Recognition. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(2), February 2015.

[95] Umapada Pal, Ramachandran Jayadevan, and Nabin Sharma. Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques. *ACM Transactions on Asian Language Information Processing*, 11(1):1–35, March 2012.

[96] U. Pal and B.B. Chaudhuri. Indian script character recognition: a survey. *Pattern Recognition*, 37(9):1887–1899, September 2004.

[97] V.K Govindan and A.P Shivaprasad. Character recognition — A review. *Pattern Recognition*, 23(7):671–683, January 1990.

[98] M. H. Glauberman. Character Recognition for Business Machines. *Electronics*, pages 132–136, February 1956.

[99] Amir Ghaderi and Vassilis Athitsos. Selective unsupervised feature learning with Convolutional Neural Network (S-CNN). In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2486–2490. IEEE, December 2016.

[100] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[101] Adam Coates, Honglak Lee, and Andrew Y. Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA.

[102] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, and Andrew Y. Ng. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In *2011 International Conference on Document Analysis and Recognition*, pages 440–445. IEEE, September 2011.

[103] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 766–774. Curran Associates, Inc., 2014.

[104] Marcal Rusinol, David Aldavert, Ricardo Toledo, and Josep Llados. Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method. In *2011 International Conference on Document Analysis and Recognition*, pages 63–67. IEEE, September 2011.

[105] Vladislavs Dovgalecs, Alexandre Burnett, Pierrick Tranouez, Stephane Nicolas, and Laurent Heutte. Spot It! Finding Words and Patterns in Historical Documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1039–1043. IEEE, August 2013.

[106] Yunxue Shao, Chunheng Wang, and Baihua Xiao. Fast self-generation voting for handwritten Chinese character recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(4):413–424, December 2013.

[107] Praneeth Shishtla, V. Surya Ganesh, Sethuramalingam Subramaniam, and Vasudeva Varma. A language-independent transliteration schema using character aligned models at NEWS 2009. In *NEWS '09 Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, page 40. Association for Computational Linguistics, 2009.

[108] Nasreen AbdulJaleel and Leah S. Larkey. English to Arabic Transliteration for Information Retrieval: A Statistical Approach. February 2004.

[109] Andrew Finch and Eiichiro Sumita. Transliteration using a Phrase-based Statistical Machine Translation System to Re-score the Output of a Joint Multigram Model. In *Proceedings of the 2010 Named Entities Workshop*, pages 48–52, Uppsala, Sweden, July 2010.

[110] Lauma Pretkalnina, Peteris Paikens, Normunds Gruzitis, Laura Rituma, and A Spektors. Making Historical Latvian Texts More Intelligible to Contemporary Readers. In *Proceedings of the LREC Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects*, May 2012.

[111] Thomas M. Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. High-Performance OCR for Printed English and Fraktur Using LSTM Networks. In *2013 12th International Conference on Document Analysis and Recognition*, pages 683–687. IEEE, August 2013.

[112] Adnan Ul-Hasan and Thomas M. Breuel. Can we build language-independent OCR using LSTM networks? In *MOCR '13 Proceedings of the 4th International Workshop on Multilingual OCR*, page 1. ACM Press, 2013.

[113] Fotini Simistira, Adnan Ul-Hassan, Vassilis Papavassiliou, Basilis Gatos, Vassilis Katsouros, and Marcus Liwicki. Recognition of historical Greek polytonic scripts using LSTM networks. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 766–770. IEEE, August 2015.

[114] B. Gatos, K. Ntirogiannis, and I. Pratikakis. DIBCO 2009: document image binarization contest. *IJDAR*, 14(1):35–44, March 2011.

[115] Friedrich M. Wahl, Kwan Y. Wong, and Richard G. Casey. Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, 20(4):375–390, December 1982.

[116] EFEO. Khmer Manuscript - Recherche - http://khmermanuscripts.efeo.fr/.

[117] Dona Valy, Michel Verleysen, Sophea Chhun, and Jean-Christophe Burie. A New Khmer Palm Leaf Manuscript Dataset for Document Analysis and Recognition – SleukRith Set. In *4th International Workshop on Historical Document Imaging and Processing*, Kyoto, Japan, November 2017.

[118] Dona Valy, Michel Verleysen, and Kimheng Sok. Line Segmentation Approach for Ancient Palm Leaf Manuscripts using Competitive Learning Algorithm. In *15th International Conference on Frontiers in Handwriting Recognition 2016*, Shenzhen, China, October 2016.

[119] Elisa H. Barney Smith. An analysis of binarization ground truthing. In *DAS '10 Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 27–34. ACM Press, 2010.

[120] Elisa H. Barney Smith and Chang An. Effect of "Ground Truth" on Image Binarization. In *DAS '12 Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*, pages 250–254. IEEE, March 2012.

[121] Nikolaos Stamatopoulos, Basilis Gatos, Georgios Louloudis, Umapada Pal, and Alireza Alaei. ICDAR 2013 Handwriting Segmentation Contest. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1402–1406. IEEE, August 2013.

[122] I.T. Phillips and A.K. Chhabra. Empirical performance evaluation of graphics recognition systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):849–870, September 1999.