Journal of *Imaging*

MDPI

*Article*

# Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia

**Made Windu Antara Kesiman [1,2,*], Dona Valy [3,4], Jean-Christophe Burie [1], Erick Paulus [5], Mira Suryani [5], Setiawan Hadi [5], Michel Verleysen [2], Sophea Chhun [4] and Jean-Marc Ogier [1]**

1 Laboratoire Informatique Image Interaction (L3i), Université de La Rochelle, 17042 La Rochelle, France; jean-christophe.burie@univ-lr.fr (J.-C.B.); jean-marc.ogier@univ-lr.fr (J.-M.O.)
2 Laboratory of Cultural Informatics (LCI), Universitas Pendidikan Ganesha, Singaraja, Bali 81116, Indonesia; michel.verleysen@uclouvain.be
3 Institute of Information and Communication Technologies, Electronic, and Applied Mathematics (ICTEAM), Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium; dona.valy@student.uclouvain.be
4 Department of Information and Communication Engineering, Institute of Technology of Cambodia, Phnom Penh, Cambodia; sophea.chhun@itc.edu.kh
5 Department of Computer Science, Universitas Padjadjaran, Bandung 45363, Indonesia; erick_paulus@yahoo.com (E.P.); mira.suryani@unpad.ac.id (M.S.); setiawanhadi@unpad.ac.id (S.H.)
* Correspondence: made_windu_antara.kesiman@univ-lr.fr

Received: 15 December 2017; Accepted: 18 February 2018; Published: 22 February 2018

**Abstract:** This paper presents a comprehensive test of the principal tasks in document image analysis (DIA), starting with binarization, text line segmentation, and isolated character/glyph recognition, and continuing on to word recognition and transliteration for a new and challenging collection of palm leaf manuscripts from Southeast Asia. This research presents and is performed on a complete dataset collection of Southeast Asian palm leaf manuscripts. It contains three different scripts: Khmer script from Cambodia, and Balinese script and Sundanese script from Indonesia. The binarization task is evaluated on many methods up to the latest in some binarization competitions. The seam carving method is evaluated for the text line segmentation task, compared to a recently new text line segmentation method for palm leaf manuscripts. For the isolated character/glyph recognition task, the evaluation is reported from the handcrafted feature extraction method, the neural network with unsupervised learning feature, and the Convolutional Neural Network (CNN) based method. Finally, the Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) based method is used to analyze the word recognition and transliteration task for the palm leaf manuscripts. The results from all experiments provide the latest findings and a quantitative benchmark for palm leaf manuscripts analysis for researchers in the DIA community.

**Keywords:** document image analysis; binarization; character recognition; text line segmentation; word recognition; transliteration; palm leaf manuscript; dataset; benchmark; experimental test

## 1. Introduction

Since the world entered the digital age in the early 20th century, the need for a document image analysis (DIA) system is increasing. This is due to the dramatic increase in efforts to digitize the various types of document collections available, especially the ancient documents of historical relics found in various parts of the world. Some very interesting projects on a wide variety of heritage document collections can be mentioned here: for example, the tranScriptorium project (http://transcriptorium.eu/) [1]; the READ (Recognition and Enrichment of Archival Documents) project (https://read.transkribus.eu/) [2], which works on documents from the Middle Ages to today, and also focuses on different languages ranging from Ancient Greek to modern English; the

IAM Historical Document Database (IAM-HistDB) (http://www.fki.inf.unibe.ch/databases/iam-historical-document-database) [3], which includes handwritten historical manuscript images from the Saint Gall Database from the 9th century in Latin; the Parzival Database from the 13th century in German; the Washington Database from the 18th century in English; the Ancient Lives Project (https://www.ancientlives.org/) [4], which asks volunteers to transcribe Ancient Greek text fragments from the Oxyrhynchus Papyri collection; and many other projects.

To accelerate the process of accessing, preserving, and disseminating the contents of the heritage documents, a DIA system is needed. Besides aiming to preserve the existence of such ancient documents physically, the DIA system is expected to enable open access to the contents of the documents and provide opportunities for a wider audience to access all the important information stored in the document. DIA is the process of using various technologies to extract text, printed or handwritten, and graphics from digitized document files (http://www.cvisiontech.com/library/pdf/pdf-document/document-image-analysis.html) [5]. DIA systems generally have a major role in identifying, analyzing, extracting, structuring, and transferring document contents more quickly, effectively, and efficiently. This system is able to work semi-automatically or even fully automatically without human intervention. The DIA system is expected to save time, cost, and effort at many points in the heritage document preservation process.

However, although the DIA research develops rapidly, it is undeniable that most of the document collections used in the initial step are from developed regions such as America and European countries. The document samples from these countries are mostly written in English or old English with Latin/Roman script. 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 deal with documents from non-English-speaking areas with non-Latin scripts, such as Arabic, Chinese, and Japanese documents. During the evolution of DIA research in the last two decades, DIA researchers have proposed and achieved satisfactory solutions for many complex problems of document analysis for these types of documents. However, the DIA research challenge is ongoing. The latest challenge is documents from Asia, with new languages and more complex scripts to explore, such as Devanagari script [6], Gurmukhi script [7–10], Bangla script [11], and Malayalam script [12], and the case of multiple languages and scripts in documents from India. Optical character recognition (OCR) for Indian languages is considered more difficult in general than for European languages because of the large number of vowels, consonants, and conjuncts (combinations of vowels and consonants) [13].

This work was part of exploring DIA research for a palm leaf manuscripts collection from Southeast Asia. This collection offers a new challenge for DIA researchers because palm leaves are used as the writing medium and the language and script have never been analyzed before. In this paper, we did a comprehensive benchmark experimental test of some principal tasks in the DIA system, starting with binarization, text line segmentation, isolated character/glyph recognition, word recognition, and transliteration. To the best of our knowledge, this work is the first comprehensive study of the DIA researchers' community and the first to perform a complete series of experimental benchmarking analyses of palm leaf manuscripts. The results of this research will be very useful in accelerating, evaluating, and improving the performance of existing DIA systems for a new type of document.

This paper is organized as follow. Section 2 gives a brief description of the palm leaf manuscripts collection from Southeast Asia, especially the Khmer palm leaf manuscript corpus from Cambodia and two palm leaf manuscript corpuses, the Balinese and Sundanese manuscripts from Indonesia. The challenges of DIA for this manuscript corpus are also presented in this section. Section 3 describes the DIA tasks that need to be developed for the palm leaf manuscript collections, followed by a description of the methods investigated for those tasks. The datasets and evaluation methods for each DIA task used in the experimental studies for this work are presented in Section 4. Section 5 reports and analyzes the detailed results of the experiments. Finally, conclusions are given in Section 6.

## 2. Palm Leaf Manuscripts from Southeast Asia

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 (form of leather made of processed sheepskin or calfskin) (http://www.casepaper.com/company/paper-history) [14], copper and bronze plates, bamboo, palm leaves, and other materials [15]. 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 bamboo and palm trees are easily found in Asia, both types of materials were the first choice of writing material in Asia. In Southeast Asia, most ancient manuscripts were written on palm leaves. For example, in Cambodia, palm leaves have been used as a writing material dating back to the first appearance of Buddhism in the country. In Thailand, dried palm leaves have also been used as one of the most popular written documents for over 500 years [16]. 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*. The existence of ancient palm leaf manuscripts in Southeast Asia is very important both in terms of the quantity and variety of historical contents.

### *2.1. Balinese Palm Leaf Manuscripts—Collection from Bali, Indonesia*

#### 2.1.1. Corpus

Apart from the collection at the museum (Museum Gedong Kertya Singaraja and Museum Bali Denpasar), it is estimated that there are more than 50,000 *lontar* collections that are owned by private families (Figure 1). For this research, in order to obtain a large variety of manuscript images, sample images have been collected from 23 different collections, which come from five different locations (regions): two museums and three private families. They consist of 10 randomly selected collections from Museum Gedong Kertya, City of Singaraja, Regency of Buleleng, North Bali, Indonesia, four collections from manuscript collections of Museum Bali, City of Denpasar, South Bali, seven collections from a private family collection from the village of Jagaraga, Regency of Buleleng, and two other private family collections from the village of Susut, Regency of Bangli and the village of Rendang, Regency of Karangasem [17].

**Figure 1.** Balinese palm leaf manuscripts.

#### 2.1.2. Balinese Script and Language

Although the official language of Indonesia, Bahasa Indonesia, is written in the Latin script, Indonesia has many local, traditional scripts, most of which are ultimately derived from Brahmi [18]. In Bali, palm leaf manuscripts were written in the Balinese script in the Balinese language, in the ancient literary texts composed in the old Javanese language of Kawi and Sanskrit. Balinese language is a Malayo-Polynesian language spoken by more than 3 million people, mainly in Bali, Indonesia (www.omniglot.com/writing/balinese.htm) [19]. Balinese is the native language of the people of Bali, known locally as Basa Bali [18]. The alphabet and numbers of Balinese script are composed of ±100 character classes including consonants, vowels, and some other special compound characters. According to the Unicode Standard 9.0, the Balinese script actually has the Unicode table from 1B00 to 1B7F.

### 2.2. Khmer Palm Leaf Manuscripts—Collection from Cambodia

#### 2.2.1. Corpus

In Cambodia, Khmer palm leaf manuscripts (Figure 2) 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 (http://khmermanuscripts.efeo.fr) [20] of microfilm images of hundreds 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 [21].

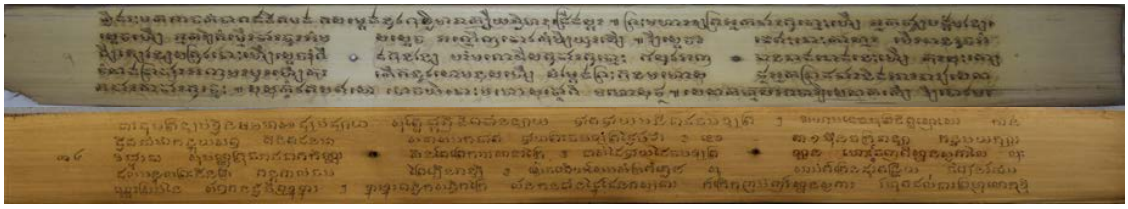

**Figure 2.** Khmer palm leaf manuscript.

#### 2.2.2. Khmer Script and Language

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.

### 2.3. Sundanese Palm Leaf Manuscripts—Collection from West Java, Indonesia

#### 2.3.1. Corpus

The collection of Sundanese palm leaf manuscripts (Figure 3) comes from Situs Kabuyutan Ciburuy, Garut, West Java, Indonesia. The Kabuyutan Ciburuy is a complex 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 15th 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 × 10–15 cm in width [22].

#### 2.3.2. Sundanese Script and Language

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.

**Figure 3.** Sundanese palm leaf manuscript.

### *2.4. Challenges of Document Image Analysis for Palm Leaf Manuscripts*

There are two main technical challenges to assessing palm leaf manuscripts in a DIA system. The first challenge is the physical condition of the palm leaf manuscript, which will strongly influence the quality of the document images captured. 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 the images [13]. Nowadays, due to the specific characteristics of the physical support of the manuscripts, the development of DIA methods for palm leaf manuscripts in order to extract relevant information is considered a new research problem in handwritten document analysis. Ancient palm leaf manuscripts contain artifacts due to aging, foxing, yellowing, strain, local shading effects, low intensity variations or poor contrast, random noises, discolored parts, fading, and other types of degradation.

The second challenge is the complexity of the script. The Southeast Asian manuscripts with different scripts and languages provide real challenges for document analysis methods, not only because of the different forms of characters in the script, but also because the writing style of each script (e.g., how to join or separate a character in a text line) differs. It ranges widely from a binarization process [23–25], text line segmentation [26,27], and character and text recognition tasks [25,28,29], to the word spotting methods [30].

In the domain of DIA, handwritten character and text 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 scripts. However, the development of handwritten character and text recognition methods for other various Asian scripts presents many issues. In the OCR task and development for palm leaf manuscripts from Southeast Asia, several 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, overlaps, and interconnection of the neighboring characters further complicate the OCR system [31]. One of the main problems faced when dealing with segmented handwritten character recognition is the ambiguity and illegibility of the characters [32]. These characteristics provide suitable conditions to test and evaluate the robustness of feature extraction methods that were proposed for character recognition.

## 3. Document Image Analysis Tasks and Investigated Methods

Heritage document preservation is not just about converting physical documents into document images. With many physical documents being digitized and stored in large document databases, and then sent and received via digital machines, the interest and demand grew to require more functionalities than simply viewing and print the images [33]. Further treatment is required before the collection of document images can be explored more extensively. For example, a more specific research field needed to be developed to add machine capabilities for extracting information from these images, reading text on a document page, finding sentences, and locating paragraphs, lines, words, and symbols on a diagram [33].

In this work, the methods for each DIA task were investigated for palm leaf manuscripts. The binarization task is evaluated using the latest methods from binarization competitions. The seam

carving method is evaluated for the text line segmentation task, compared to a recent text line segmentation method for palm leaf manuscripts [27]. For the isolated character/glyph recognition task, the evaluation is reported from the handcrafted feature extraction method, the neural network with unsupervised learning feature to the CNN based method. Finally, the RNN-LSTM based method is used to analyze the word recognition and transliteration task for palm leaf manuscripts.

### 3.1. Binarization

Binarization is widely applied as the first pre-processing step in image document analysis [34]. Binarization is a common starting point for document image analysis and converts gray image values into binary representation for background and foreground, or, more specifically, text and non-text, which is then fed into further document processing tasks such as text line segmentation and optical character recognition. The performance of binarization techniques has a great impact and directly affects the performance of the recognition task [35]. Non-optimal binarization methods produce unrecognizable characters with noise [16]. Many binarization methods have been reported. These methods have been tested and evaluated on different types of document collections. Based on the choice of the thresholding value, binarization methods can generally be divided into two types, global binarization and local adaptive binarization [16]. Some surveys and comparative studies of the performance of several binarization methods have been reported [35,36]. A binarization method that performs well for one document collection may not necessarily be applied to another document collection with the same performance [34]. 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 the historical archive documents [36].

In this work, we compared several alternative binarization algorithms for palm leaf manuscripts. We tested and evaluated some well-known standard binarization methods, and some binarization methods that are experimentally promising for historical archive documents, though not specifically for images of palm leaf manuscripts. We also tested the binarization methods from the Document Image Binarization Competition (DIBCO) competition [37,38], for example Howe's method [39] and the ones from the International Conference on Frontiers in Handwriting Recognition (ICFHR) competition (amadi.univ-lr.fr/ICFHR2016_Contest) [25,40].

#### 3.1.1. Global Thresholding

Global thresholding is the simplest technique and the most conventional approach for binarization [34,41]. A single threshold value was calculated from the 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 promising optimal binarization results [36]. 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 in the foreground and background colors on low-quality document images gives unsatisfactory results. It is difficult to choose one fixed threshold value that is adaptable for all images [36,42].

Otsu's method is a very popular global binarization technique [34,41]. 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 [34]. Otsu's method is implemented as a standard binarization technique in a built-in Matlab function called *graythresh* (https://fr.mathworks.com/help/images/ref/graythresh.html) [43].

#### 3.1.2. Local Adaptive Binarization

To overcome the weakness of the global binarization technique, many local adaptive binarization techniques were proposed, for example Niblack's method [34,36,41,42,44], Sauvola's method [34,36,41,42,44,45], Wolf's method [42,44,46], NICK method [44], and the Rais method [34]. The threshold value in local adaptive binarization technique is calculated in each smaller local image

area, region, or window. Niblack's method proposed a 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 results for many document collections. However, there is still a drawback to this method. It was found that Niblack's method works optimally only on the text region, but is not well suited for large non-text regions of an image. The absence of text in local areas forces Niblack's method to detect noise as text. The suitable window size should be chosen based on the character and stroke size, which may vary for each image. Many other local adaptive binarization techniques were proposed to improve the performance of the basic Niblack 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, large variations, and uneven illumination. The improvement over Niblack's method is in the use of adaptive contribution of standard deviation in determining the local threshold on the gray values of text and non-text pixels. Sauvola's method processes the image in $N \times N$ 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 to Niblack's method are NICK method and the Rais method. NICK method proposes a threshold computation derived from the basic Niblack's method and the Rais method proposes an optimal size of window for the local binarization.

#### 3.1.3. Training-Based Binarization

The top two proposed methods in the Binarization Challenge for the ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts are training-based binarization methods [25]. The best method in this competition employs a Fully Convolutional Network (FCN). It takes a color subimage as input and outputs the probability that each pixel in the sub-image is part of the foreground. The FCN is pre-trained on normal handwritten document images with automatically generated "ground truth" binarizations (using the method of Wolf et al. [46]). 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. Consequently, 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.

The second-best method uses two neural network classifiers, $C_1$ and $C_2$, to classify each pixel as background or not. Two binarized images, $B_1$ and $B_2$, are generated in this step. $C_1$ is a rough classifier that tries to detect all the foreground pixels, while probably making mistakes for some background pixels. $C_2$ is an accurate classifier that should not classify a background pixel as a foreground pixel but probably misses some foreground pixels. Secondly, these two binary images are joined to get the final classification result.

### *3.2. Text Line Segmentation*

Text line segmentation is a crucial pre-processing step in most DIA pipelines. The task aims at extracting and separating text regions into individual lines. Most line segmentation approaches in the literature require that the input image be binarized. However, due to the degradation and noise often found in historical documents such as palm leaf manuscripts, the binarization task is not able to produce good enough results (see Section 5.1). In this paper, we investigate two line segmentation methods that are independent of the binarization task. These approaches work directly on color/grayscale images.

#### 3.2.1. Seam Carving Method

Arvanitopoulos and Süsstrunk [47] 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, accumulating energy. The path with the minimum cumulative energy is then chosen.

#### 3.2.2. Adaptive Path Finding Method

This approach was proposed by Valy et al. [27]. The method takes as input a grayscale image of a document page. 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 that come from noise and artifacts) is used to create a stroke map. Using column-wise projection profiles on the output map, estimated number and medial positions of 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 the foreground text (intensity difference cost function $D$) and another one favoring the path that stays close to the estimated medial lines (vertical distance cost function $V$). Figure 4 illustrates an example of an optimal path.

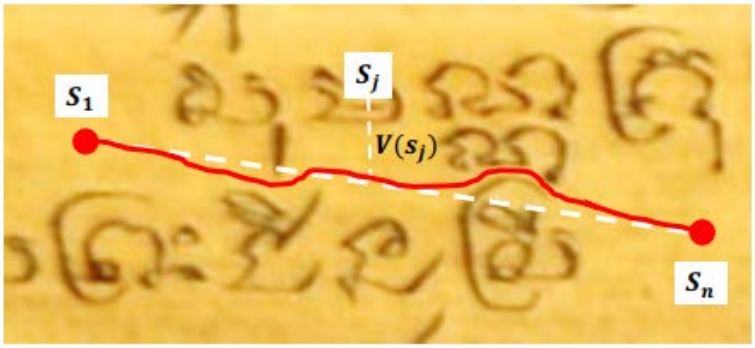


**Figure 4.** An example of an optimal path going from start state $S_1$ to goal state $S_n$.

### *3.3. Isolated Character/Glyph Recognition*

In a DIA system, word or text recognition tasks are generally categorized into two different approaches: segmentation-based and segmentation-free methods. In segmentation-based methods, the isolated character recognition task is a very important process [9]. A proper feature extraction and a correct classifier selection can increase the recognition rate [48]. Although many methods for isolated character recognition have been developed and tested, especially for Latin-based scripts and alphabets, there is still a need for in-depth evaluation of those methods as applied to various other scripts. This includes the isolated character recognition task for many Southeast Asian scripts, and more specifically scripts that were written on ancient palm leaf manuscripts.

Previous studies on isolated character recognition in palm leaf manuscripts have already been reported, but only with the Balinese script as the benchmark dataset [28,29]. In that first work, an experimental study on feature extraction methods for character recognition of Balinese script was performed [28]. For the second work, a training-based method with neural network and unsupervised

feature learning was used to increase the recognition rate [29]. In this paper, we will conduct a broader evaluation of the robustness of the methods previously tested on Balinese script, using the other two palm leaf manuscripts with Khmer and Sundanese scripts. In the next sub-sections, we provide a brief description of the methods. For a detailed description of each method, interested readers can refer to our previous works.

#### 3.3.1. Handcrafted Feature Extraction Methods

Since the beginning of pattern recognition research, many feature extraction methods for character recognition have been presented in the literature. In our previous work [28], we investigated and evaluated 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. After evaluating the performance of those individual feature extraction methods, we found that the Histogram of Gradient (HoG) features as directional gradient-based features [9,49] (Figure 5), the Neighborhood Pixels Weights (NPW) [50] (Figure 6), the Kirsch Directional Edges [50], and Zoning [12,32,50,51] (Figure 7) give very promising results. We then proposed a new feature extraction method applying NPW on Kirsch edge images (Figure 8) and concatenated the NPW–Kirsch with two other features, HoG and Zoning method, with k-NN as the classifier.

**Figure 5.** The representation of the array of cells in HoG [28].

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 3 | 3 | | 3 | 3 | 3 |
| 3 | 2 | 2 | | 2 | 2 | 3 |
| 3 | 2 | 1 | | 1 | 2 | 3 |
| | | | P | | | |
| 3 | 2 | 1 | | 1 | 2 | 3 |
| 3 | 2 | 2 | | 2 | 2 | 3 |
| 3 | 3 | 3 | | 3 | 3 | 3 |

**Figure 6.** Neighborhood pixels for NPW features [28].

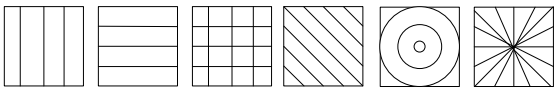

**Figure 7.** Type of Zoning (from left to right: vertical, horizontal, block, diagonal, circular, and radial zoning) [28].

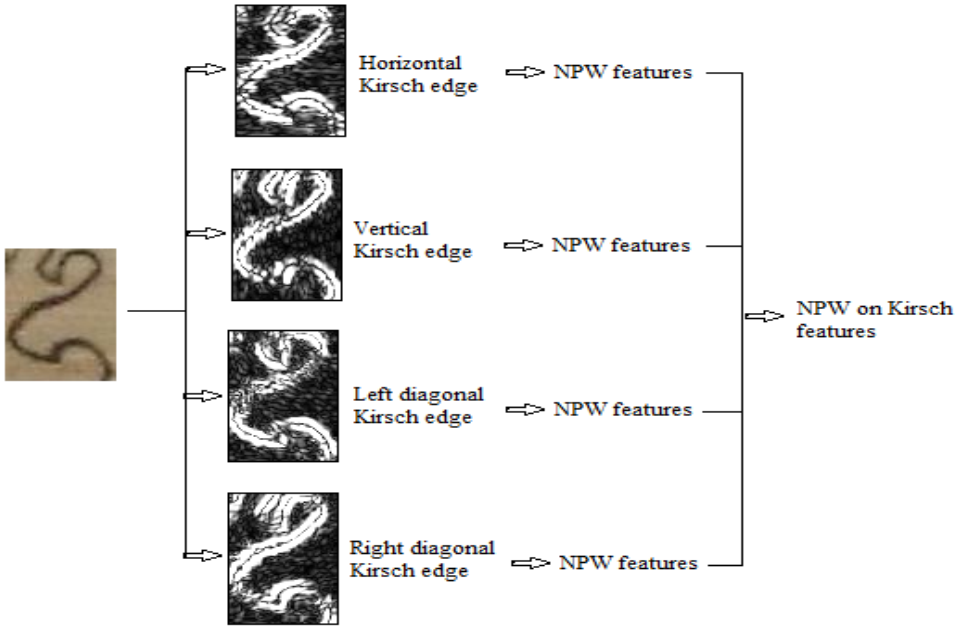


**Figure 8.** Scheme of NPW on Kirsch features [28].

### 3.3.2. Unsupervised Learning Feature and Neural Network

With the aim of improving the performance of our proposed feature extraction method, we continued our research on isolated character recognition by implementing the neural network as classifier. In this second step [29], the same combination of feature extraction methods was used and sent as the input feature vector to a single-layer neural network character recognizer. 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 9). This schema was inspired by the study of Coates et al. [52,53]. The unsupervised learning calculates the initial learning weight for the neural network training phase from the cluster centers of all feature vectors.

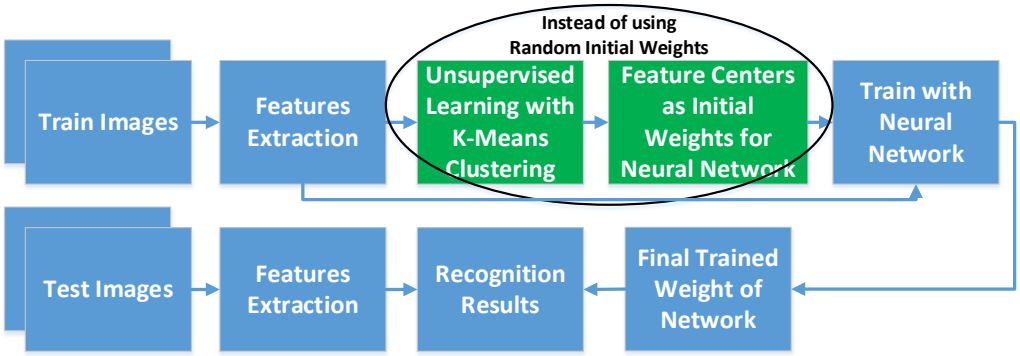


**Figure 9.** Schema of character recognizer with feature extraction method, unsupervised learning feature, and neural network [29].

### 3.3.3. Convolutional Neural Network

The multilayer convolutional neural networks (CNN) have proven very effective in areas such as image recognition and classification. In this evaluation experiment, a vanilla CNN is used. The architecture of the CNN (Figure 10) is described as follows (this architecture has also been reported in Khmer isolated character recognition baseline in [21]). The grayscale input images of isolated characters are rescaled to 48 × 48 pixels in size and 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 the ReLu function and followed by a max pooling of $2 \times 2$ blocks. The numbers of feature maps (of size $5 \times 5$) 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 (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 an Adam optimizer with a batch size of 100 and a learning rate of 0.0001.

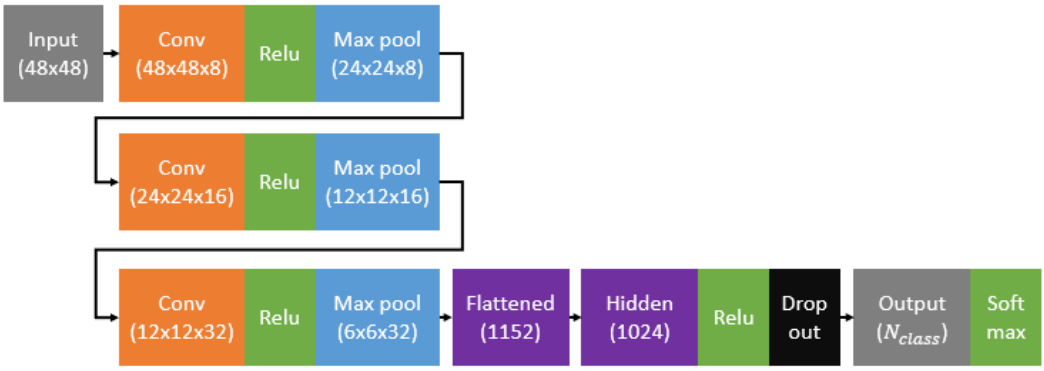


**Figure 10.** Architecture of the CNN.

### 3.4. Word Recognition and Transliteration

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. In many DIA systems, word or text recognition is the final task in the processing pipeline. However, normally in Southeast Asian script the speech sound of the syllable change is related to some certain phonological rules. In this case, an OCR system is not enough. Therefore, a transliteration system should also be developed to help transliterate the ancient scripts on these manuscripts. By definition, transliteration is defined as the process of obtaining the phonetic translation of names across languages [54]. Transliteration involves rendering a language from one writing system to another. In [54], the problem is stated formally as a sequence labeling problem from one language alphabet to another. It will help us to index and to quickly and efficiently access the content of the manuscripts. In our previous work [29], a complete scheme for segmentation-based glyph recognition and transliteration specific to Balinese palm leaf manuscripts was proposed. In this work, a segmentation-free method will be evaluated to recognize and transliterate the words from three different scripts of a palm leaf manuscript.

#### RNN/LSTM-Based Methods

From the last decade, sequence-analysis-based methods using a Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) type of learning network have been very popular among researchers in text recognition. RNN-LSTM-based method together 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 raw image pixel can be sent directly as the input to the learning network and there is no requirement to segment the training data sequence. RNN is basically an extended version of the basic feedforward neural network. In a RNN, the neurons in the hidden layer are connected to each other. RNN offers very good context-aware processing to recognize patterns in a sequence or time series. One drawback of RNN is the vanishing gradient problem. To deal with this problem, the LSTM architecture was introduced. The LSTM network adds multiplicative gates and additive feedback. Bidirectional LSTM is an LSTM

architecture with two-directional (forward and backward) context processing. LSTM architecture is widely evaluated as a generic and language-independent text recognizer [55]. In this work, the OCRopy (https://github.com/tmbdev/ocropy) [56] framework is used to test and evaluate the word recognition and transliteration tasks for the palm leaf manuscript collection. OCRopy provides the functional library of the OCR system by using RNN-LSTM architecture (http://graal.hypotheses.org/786) [57,58]. We evaluated the dataset with unidirectional LSTM and the (Bidirectional LTSM) BLSTM architecture.

## 4. Experiments: Datasets and Evaluation Methods

From the three manuscript corpuses (Khmer, Balinese, and Sundanese), the datasets for each DIA task were extracted and used in the experimental work for this research.

### *4.1. Binarization*

#### 4.1.1. Datasets

The palm leaf manuscript datasets for binarization task are presented in Table 1. For Khmer manuscripts, one ground truth binarized image is provided for each image, but for Balinese and Sundanese manuscripts, each image has two different ground truth binarized images [17,25]. The study of ground truth variability and subjectivity was reported in the previous work [24]. 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 11). A pressure-sensitive tip stylus is used to trace each text stroke by keeping the original size of the stroke width [59]. For the manuscripts from Bali, the binarized ground truth images have been created with a semi-automatic scheme [17,23–25] (Figure 12). The binarized ground truth images for Sundanese manuscripts were manually [22] generated using PixLabeler [60] (Figure 13). The training set is provided only for the Balinese dataset. We used all images of the Khmer and Sundanese corpuses as a test set because the training-based binarization method (ICFHR G1 method, see Section 5.1) was evaluated for the Khmer and Sundanese datasets by using only the pre-trained Balinese training set weighted model.

**Table 1.** Palm leaf manuscript datasets for binarization task.

| Manuscripts | Train | Test | Ground Truth | Dataset |
|---|---|---|---|---|
| Balinese | 50 pages | 50 pages | 2 × 100 pages | Extracted from AMADI_LontarSet [17,25,40] |
| Khmer | - | 46 pages | 1 × 46 pages | Extracted from EFEO [20,59] |
| Sundanese | - | 61 pages | 2 × 61 pages | Extracted from Sunda Dataset ICDAR2017 [22] |

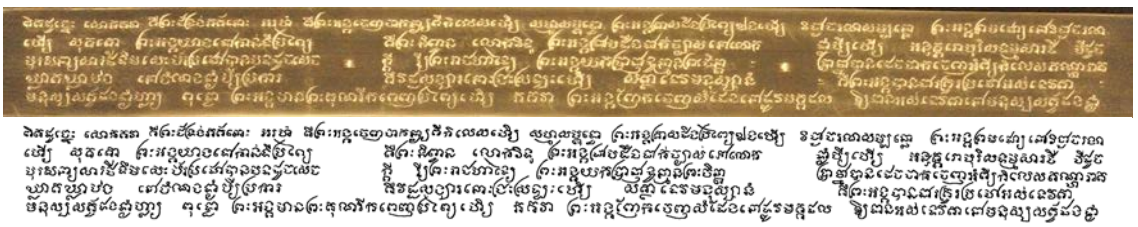

**Figure 11.** Khmer manuscript with binarized ground truth image.

**Figure 12.** Balinese manuscript with binarized ground truth image.

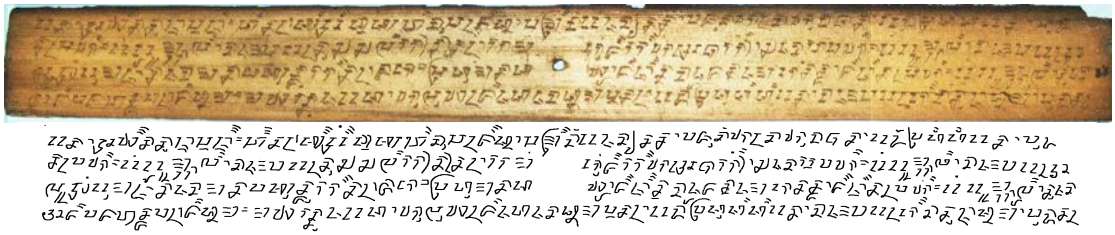

**Figure 13.** Sundanese manuscript with binarized ground truth image.

#### 4.1.2. Evaluation Method

Following our previous work [24] and the evaluation method from the ICFHR competition [25], three metrics of binarization evaluation that were used in the DIBCO 2009 contest [37] are used in the binarization task evaluation for this work. Those three metrics are F-Measure (FM) (Equation (3)), *Peak SNR* (*PSNR*) (Equation (5)), and Negative Rate Metric (NRM) (Equation (8)).

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

$$\text{Re}call = \frac{TP}{FN + TP} \times 100 \tag{1}$$

$$\text{Pr}ecision = \frac{TP}{FP + TP} \times 100 \tag{2}$$

*TP*, defined as true positive, occurs when the image pixel is labeled as foreground and the ground truth is also. *FP*, defined as false positive, occurs when the image pixel is labeled as foreground but the ground truth is labeled as background. *FN*, defined as false negative, occurs when the image pixel is labeled as background but the ground truth is labeled as foreground (Equations (1) and (2)).

$$FM = \frac{2 \times \text{Re}call \times \text{Pr}ecision}{\text{Re}call + \text{Pr}ecision} \tag{3}$$

A higher F-measure indicates a better match.

*Peak SNR (PSNR)*: PSNR is calculated from *Mean Square Error (MSE)* (Equation (4)).

$$MSE = \sum_{x=1}^{M} \sum_{y=1}^{N} \frac{(I_1(x,y) - I_2(x,y))^2}{M * N} \tag{4}$$

$$PSNR = 10 \times \log_{10}(\frac{C^2}{MSE}), \tag{5}$$

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

*Negative Rate Metric (NRM)*: NRM is defined from the negative rate of false negative ($NR_{FN}$) (Equation (6)) and the negative rate of false positive ($NR_{FP}$) (Equation (7)):

$$NR_{FN} = \frac{FN}{FN + TP} \tag{6}$$

$$NR_{FP} = \frac{FP}{FP + TN} \tag{7}$$

*TN*, defined as true negative, occurs when both the image pixel and ground truth are labeled as background. The definitions of *TP*, *FN*, and *FP* are the same as the ones given for the F-Measure.

$$NRM = \frac{NR_{FN} + NR_{FP}}{2} \tag{8}$$

A lower NRM indicates a better match.

### 4.2. Text Line Segmentation

#### 4.2.1. Datasets

The palm leaf manuscript datasets for text line segmentation task are presented in Table 2. The text line segmentation ground truth data for Balinese and Sundanese manuscripts have been generated by hand based on the binarized ground truth images [17]. For Khmer 1, a semi-automatic scheme is used [26,59]. 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 [21,27].

**Table 2.** Palm leaf manuscript datasets for text line segmentation task.

| Manuscripts | Pages | Text Lines | Dataset |
|---|---|---|---|
| Balinese 1 | 35 pages | 140 text lines | Extracted from AMADI_LontarSet [17,26,40] |
| Balinese 2 | Bali-2.1: 47 pages<br>Bali-2.2: 49 pages | 181 text lines<br>182 text lines | Extracted from AMADI_LontarSet [17] |
| Khmer 1 | 43 pages | 191 text lines | Extracted from EFEO [20,26,59] |
| Khmer 2 | 100 pages | 476 text lines | Extracted from SleukRith Set [21,27] |
| Khmer 3 | 200 pages | 971 text lines | Extracted from SleukRith Set [21] |
| Sundanese 1 | 12 pages | 46 text lines | Extracted from Sunda Dataset [26] |
| Sundanese 2 | 61 pages | 242 text lines | Extracted from Sunda Dataset [22] |

#### 4.2.2. Evaluation Method

Following our previous work [26], we use the evaluation criteria and tool provided by ICDAR2013 Handwriting Segmentation Contest [61]. First, the one-to-one (o2o) match score is computed for a region pair based on the evaluator's acceptance threshold. In our experiments, we used 90% as the acceptance threshold. Let N be the count of ground truth elements, and M the count of result elements. With the o2o score, three metrics are calculated: detection rate (DR), recognition accuracy (RA), and performance metric (FM).

### 4.3. Isolated Character/Glyph Recognition

#### 4.3.1. Datasets

The palm leaf manuscript datasets for isolated character/glyph recognition task are presented in Table 3. For the Balinese character dataset, Balinese philologists manually annotated the segment

of connected components that represented a correct character in Balinese script from the word-level binarized images that were manually annotated [11,17,20] using Aletheia (http://www.primaresearch.org/tools/Aletheia) [62,63] (Figure 14). The Sundanese character dataset was annotated manually [22] (Figure 15). 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 [21] (Figure 16).

**Table 3.** Palm leaf manuscript datasets for isolated character/glyph recognition task.

| Manuscripts | Classes | Train | Test | Dataset |
|---|---|---|---|---|
| Balinese | 133 classes | 11,710 images | 7673 images | AMADI_LontarSet [17,25,28] |
| Khmer | 111 classes | 113,206 images | 90,669 images | SleukRith Set [21] |
| Sundanese | 60 classes | 4555 images | 2816 images | Sunda Dataset [22] |

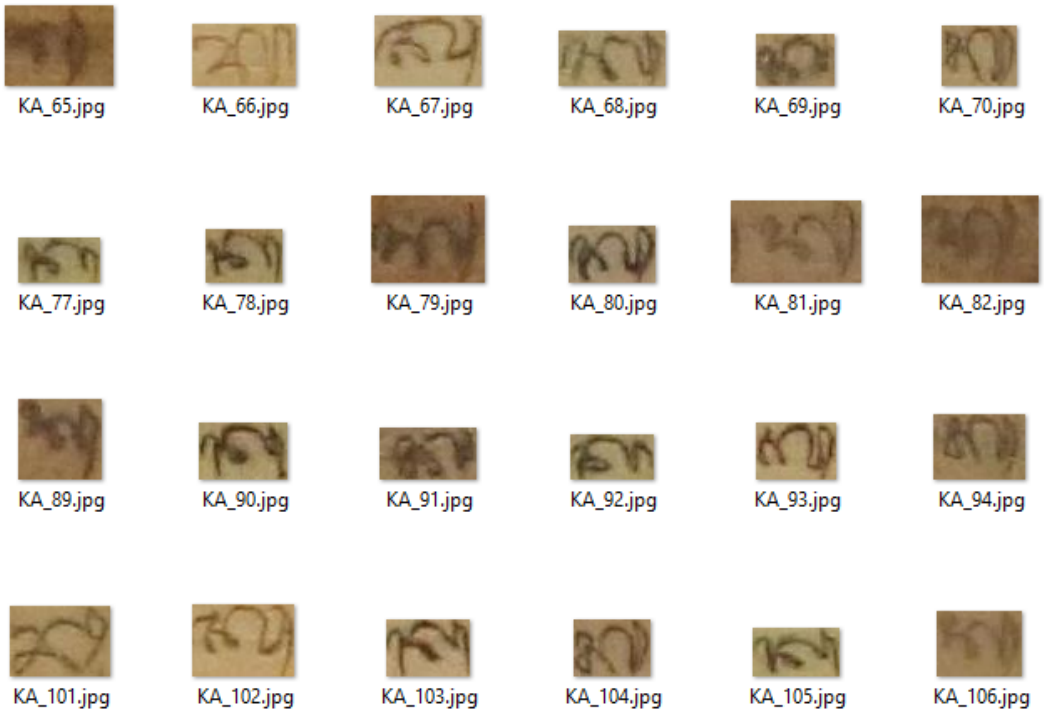


**Figure 14.** Balinese character dataset.

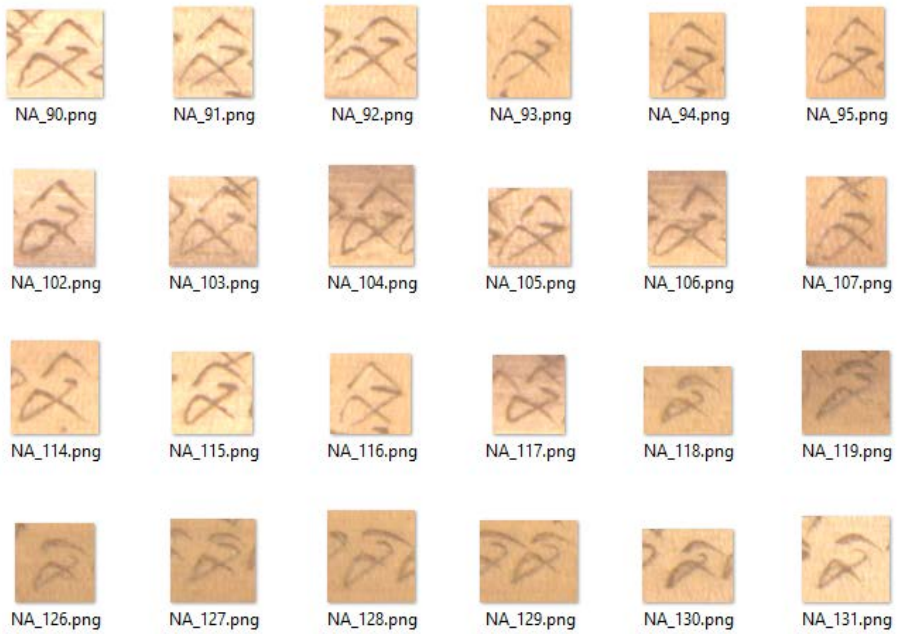


**Figure 15.** Sundanese character dataset.


**Figure 16.** Khmer character dataset.

#### 4.3.2. Evaluation Method

Following the evaluation method from the ICFHR competition [25], the recognition rate, i.e., the percentage of correctly classified samples over the test samples (C/N) is calculated, where C is the number of correctly recognized samples and N is the total number of test samples.

### *4.4. Word Recognition and Transliteration*

#### 4.4.1. Datasets

The palm leaf manuscript datasets for word recognition and transliteration task are presented in Table 4. For the Khmer dataset, all characters on the page have been annotated and grouped together

into words (Figure 17). 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 [21]. Balinese (Figure 18) and the Sundanese (Figure 19) word dataset was manually annotated using Aletheia [63].

**Table 4.** Palm leaf manuscript datasets for word recognition and transliteration tasks.

| Manuscripts | Train | Test | Text | Published |
|---|---|---|---|---|
| Balinese | 15,022 images from 130 pages | 10,475 images from 100 pages | Latin | AMADI_LontarSet [17,25] |
| Khmer | 16,333 images (part of 657 pages) | 7791 images (part of 657 pages) | Latin and Khmer | SleukRith Set [21] |
| Sundanese | 1427 images from 20 pages | 318 images from 10 pages | Latin | Sunda Dataset [22] |


**Figure 17.** Khmer word dataset.


**Figure 18.** Balinese word dataset.

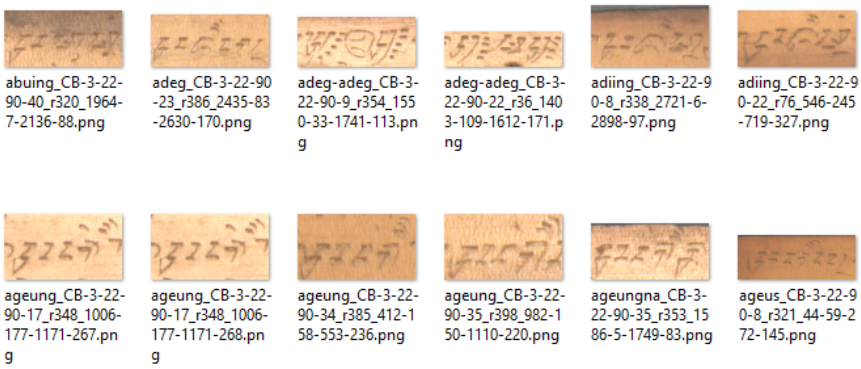


**Figure 19.** Sundanese word dataset.

#### 4.4.2. Evaluation Method

The error rate is defined by edit distances between ground truth and recognizer output and is computed using the provided OCRopy function ocropus-errs (https://github.com/tmbdev/ocropy/blob/master/ocropus-errs) [56].

## 5. Experimental Results and Discussion

In this section, the performance of each method for the DIA tasks on palm leaf manuscript collections is presented.

### *5.1. Binarization*

The experimental results for the binarization task are presented in Table 5. These results show that the performance of all methods on each dataset is still quite low. Most of the methods achieve less than a 50% FM score. This means that palm leaf manuscripts are still an open challenge for the binarization task. The different parameter values for the local adaptive binarization methods show significant improvement in performance, but still give unsatisfactory results. In these experiments, the ICFHR G1 method was evaluated for the Khmer and Sundanese datasets 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 20), ICFHR G1 method gives the highest FM score for Khmer manuscripts (Figure 21), and ICFHR G2 gives the highest FM score for Balinese manuscripts (Figure 22). However, visually, there are still many broken and unrecognizable characters/glyphs, and noise is detected in the images.

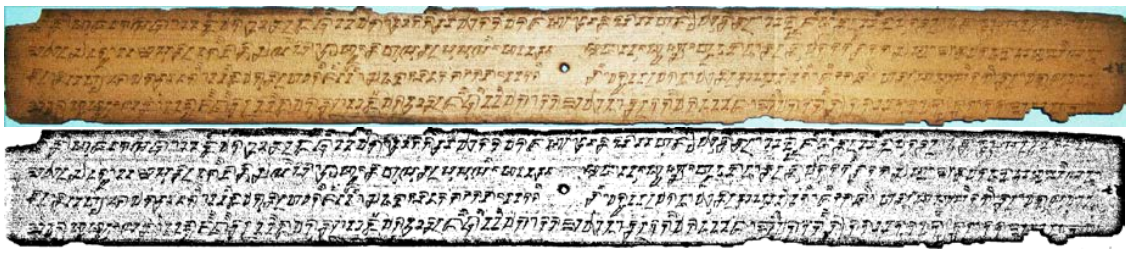

**Figure 20.** Binarization of Sundanese manuscript with Niblack's method.

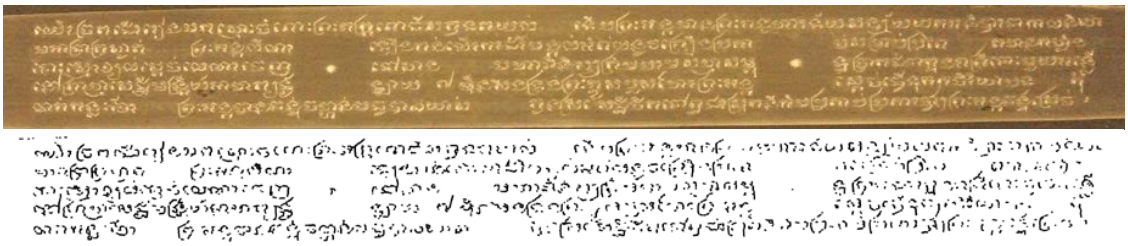

**Figure 21.** Binarization of Khmer manuscript with ICFHR G1 method.

**Table 5.** 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 [34,41] | Otsu from gray image Using Matlab graythresh [43] | Balinese | 18.98178 | 0.398894 | 5.019868 |
| | | Khmer | 23.92159 | 0.313062 | 7.387765 |
| | | Sundanese | 23.70566 | 0.326681 | 9.998433 |
| OtsuRed [34,41] | 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 [34,36,41,42,44,45] | $window = 50$, $k = 0.5$, $R = 128$ | Balinese | 13.20997 | 0.462312 | 27.69732 |
| | | Khmer | 44.73579 | 0.268527 | 26.06089 |
| | | Sundanese | 6.190919 | 0.479984 | 24.78595 |
| Sauvola2 [34,36,41,42,44,45] | $window = 50$, $k = 0.2$, $R = 128$ | Balinese | 40.18596 | 0.274551 | 25.0988 |
| | | Khmer | 47.55924 | 0.155722 | 21.96846 |
| | | Sundanese | 43.04994 | 0.299694 | 23.65228 |
| Sauvola3 [34,36,41,42,44,45] | $window = 50$, $k = 0.0$, $R = 128$ | Balinese | 35.38635 | 0.165839 | 17.05408 |
| | | Khmer | 30.5562 | 0.190081 | 12.78953 |
| | | Sundanese | 40.29642 | 0.181465 | 16.25056 |
| Niblack [34,36,41,42,44] | $window = 50$, $k = -0.2$ | Balinese | 41.55696 | 0.175795 | 21.24452 |
| | | Khmer | 38.01222 | 0.160807 | 16.84153 |
| | | Sundanese | 46.79678 | 0.195015 | 20.31759 |
| Niblack2 [34,36,41,42,44] | $window = 50$, $k = 0.0$ | Balinese | 35.38635 | 0.165839 | 17.05408 |
| | | Khmer | 30.5562 | 0.190081 | 12.78953 |
| | | Sundanese | 40.29642 | 0.181465 | 16.25056 |
| NICK [44] | $window = 50$, $k = -0.2$ | Balinese | 37.85919 | 0.328327 | 27.59038 |
| | | Khmer | 51.2578 | 0.176003 | 24.51998 |
| | | Sundanese | 29.5918 | 0.390431 | 24.26187 |
| Rais [34] | $window = 50$ | Balinese | 34.46977 | 0.171096 | 16.84049 |
| | | Khmer | 31.59138 | 0.187948 | 13.52816 |
| | | Sundanese | 40.65458 | 0.177016 | 16.35472 |
| Wolf [42,44] | $window = 50$, $k = 0.5$ | Balinese | 27.94817 | 0.392937 | 27.1625 |
| | | Khmer | 46.78589 | 0.23739 | 25.1946 |
| | | Sundanese | 42.40799 | 0.299157 | 23.61075 |
| Howe1 [39] | Default values [39] | Balinese | 44.70123 | 0.267627 | 28.35427 |
| | | Khmer | 40.20485 | 0.280604 | 25.59887 |
| | | Sundanese | 45.90779 | 0.235175 | 21.90439 |
| Howe2 [39] | Default values | Balinese | 40.5555 | 0.273994 | 28.02874 |
| | | Khmer | 32.35603 | 0.294016 | 25.96965 |
| | | Sundanese | 35.35973 | 0.274865 | 22.36583 |
| Howe3 [39] | Default values | Balinese | 42.15377 | 0.304962 | 28.38466 |
| | | Khmer | 30.7186 | 0.382087 | 26.36983 |
| | | Sundanese | 25.77321 | 0.350349 | 23.66912 |
| Howe4 [39] | 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. [25] | 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. [25] | Balinese | 68.76 | 0.13 | 33.39 |
| | | Khmer | - | - | - |
| | | Sundanese | - | - | - |
| ICFHR G3 | See ref. [25] | Balinese | 52.20 | 0.18 | 26.92 |
| | | Khmer | - | - | - |
| | | Sundanese | - | - | - |
| ICFHR G4 | See ref. [25] | Balinese | 58.57 | 0.17 | 29.98 |
| | | Khmer | - | - | - |
| | | Sundanese | - | - | - |

**Figure 22.** Binarization of Balinese manuscript with ICFHR G2 method.

### 5.2. Text Line Segmentation

The experimental results for text line segmentation task are presented in Table 6. According to these results, both methods perform sufficiently well for most datasets, except Khmer 1 (Figures 23–25). 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 achieves 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 adjacent 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 the seam borders between two neighboring lines, so they can be considered 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 are important. 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.

**Table 6.** Experimental results for 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 [47] | 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 [27] | 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 23.** Text line segmentation of Balinese manuscript with the Seam Carving method (green) and Adaptive Path Finding (red).

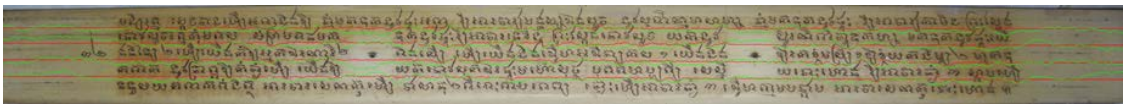

**Figure 24.** Text line segmentation of Khmer manuscript with the Seam Carving method (green) and Adaptive Path Finding (red).

**Figure 25.** Text line segmentation of Sundanese manuscript with the Seam Carving method (green) and Adaptive Path Finding (red).

### 5.3. Isolated Character/Glyph Recognition

The experimental results for isolated character/glyph recognition task are presented in Table 7. 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. According to these results, the handcrafted feature extraction combination of HoG-NPW-Kirsch-Zoning is a proper choice resulting in a good recognition rate for Balinese and Khmer characters/glyphs. The CNN methods also show satisfactory results, but the differences in recognition rates are not too significant with the handcrafted feature combinations. The unbalanced number of image samples for each character class means the CNN method did not perform optimally. For the Sundanese dataset, the handcrafted feature with NN slightly outperformed the CNN method. The UFL method slightly increased the recognition rate of the pure NN method for the Khmer and Balinese datasets.

**Table 7.** Experimental results for isolated character/glyph recognition tasks (in % recognition rate).

| Methods | Balinese | Khmer | Sundanese |
|---|---|---|---|
| Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with k-NN [28] | 85.16 | - | 72.91 |
| Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with NN [29] | 85.51 | 92.15 | 79.69 |
| Handcrafted Feature (HoG-NPW-Kirsch-Zoning) with UFL + NN [29] | 85.63 | 92.44 | 79.33 |
| CNN 1 [28] | 84.31 | - | - |
| CNN 2 | 85.39 | 93.96 | 79.05 |
| ICFHR G1: VCMF [25] | 87.44 | - | - |
| ICFHR G1: VMQDF [25] | 88.39 | - | - |
| ICFHR G3 [25] | 77.83 | - | - |
| ICFHR G5 [25] | 77.70 | - | - |

### 5.4. Word Recognition and Transliteration

The experimental results for word recognition and transliteration task are presented in Table 8. The error rates for word recognition and transliteration tests set on each training model iteration are shown in Figures 26–28. The LSTM-based architecture of OCRopy seems very promising in terms of

recognizing and directly transliterating Balinese words. For the Khmer and Sundanese datasets, the LSTM architecture seems to struggle to learn the training data. More synthetic data training with a more frequent word should be generated in order to support the training process. For the Balinese dataset, a sequence depth of 100 pixels with a neuron size of 200 gives a better result for both LSTM and BLTSM architecture. Most of the Southeast Asian scripts are syllabic scripts. One character/glyph in these scripts represents a syllable, with a sequence of letters in 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.

**Table 8.** Experimental results for word recognition and transliteration tasks (in % error rate for test).

| Methods (with OCRopy [56] Framework) | Balinese | Khmer | Sundanese |
|---|---|---|---|
| BLSTM 1 (seq_depth 60, neuron size 100) | 43.13 | Latin text: 73.76<br>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 | - | - |

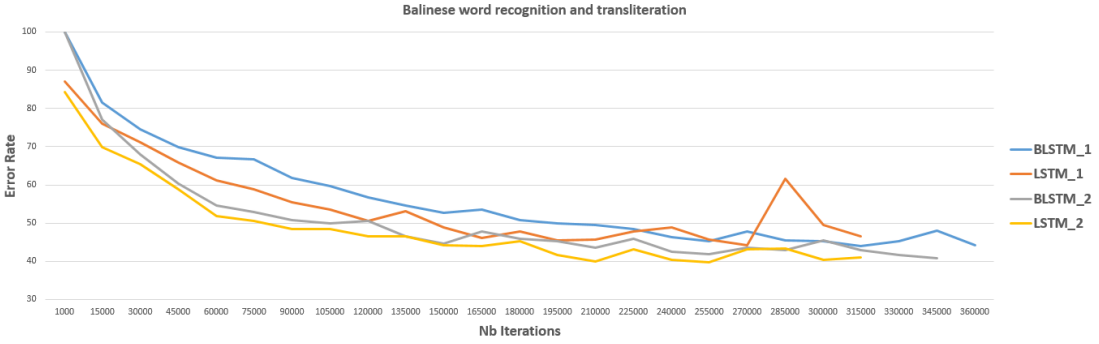


**Figure 26.** Error rate for Balinese word recognition and transliteration test set.

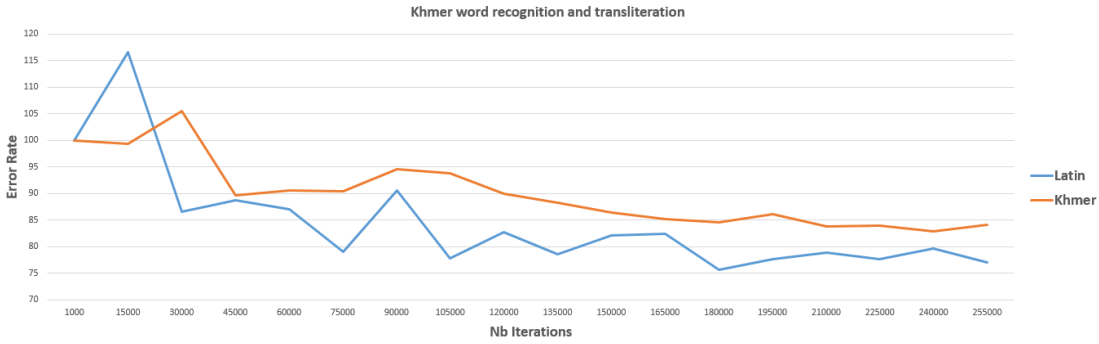


**Figure 27.** Error rate for Khmer word recognition and transliteration test set.

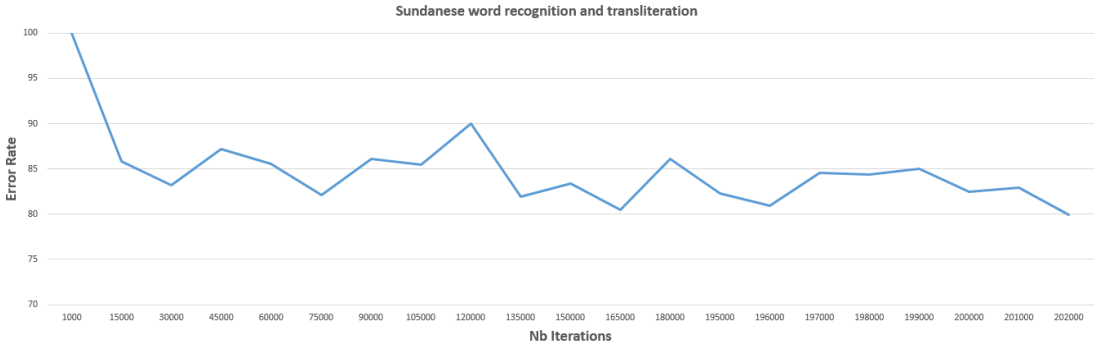


**Figure 28.** Error rate for Sundanese word recognition and transliteration test set.

## 6. Conclusions and Future Work

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, is presented. The results from all experiments provide the latest findings and a quantitative benchmark of palm leaf manuscripts analysis for researchers in the DIA community. Binarizing the palm leaf manuscript images seems very challenging. 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. The problem of inadequate training data also influences glyph recognition and word transliteration. The unbalanced number of image samples for each character class means the CNN methods did not perform optimally in glyph recognition. The differences in the recognition rates of the CNN methods are not too significant with the handcrafted feature combinations. For future work, more synthetic data training for palm leaf manuscript images should be generated in order to support the training process. Especially for the word transliteration task, more synthetic data training with a more frequent word should be generated in order to improve the training process. Many examples of glyph-to-syllable association should be synthetically generated to transliterate syllabic scripts from Southeast Asia. The special characteristics and challenges posed by the palm leaf manuscript collections will require a thorough adaptation of the DIA system. Some specific adjustments need to be applied to the DIA methods for other types of documents. The adaptation of a DIA for palm leaf manuscripts is not unique and is not universal for all types of problem 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.

**Acknowledgments:** The authors would like to thank Museum Gedong Kertya, Museum Bali, Undang Ahmad Darsa, the philologists from Sundanese Centre Studies of Universitas Padjadjaran, the Situs Kabuyutan Ciburuy Garut, all families in Bali, Indonesia, the EFEO team, the Buddhist Institute, and the National Library in Cambodia for providing us with samples of palm leaf manuscripts. We also thank the students from the Department of Informatics Education and the Department of Balinese Literature, University of Pendidikan Ganesha, the Institute of Technology of Cambodia, and the National Institute of Post, Telecommunication and ICT for helping us with the ground truthing process for this research project. This work is supported by the DIKTI BPPLN Indonesian Scholarship Program, the STIC Asia Program implemented by the French Ministry of Foreign Affairs and International Development (MAEDI), and ARES-CCD (program AI 2014-2019) under the funding of Belgian university cooperation, and DRPMI Universitas Padjadjaran, DIKTI International Collaboration and Publication grant 2017.

**Author Contributions:** The Balinese dataset was prepared by Made Windu Antara Kesiman. The Khmer dataset was prepared by Dona Valy and Sophea Chhun. The Sundanese dataset was prepared by Erick Paulus, Mira Suryani, and Setiawan Hadi. Jean-Christophe Burie, Michel Verleysen, and Jean-Marc Ogier contributed to designing a ground truth validation protocol. Made Windu Antara Kesiman and Dona Valy conceived, designed,

and performed the experiments. Made Windu Antara Kesiman, Dona Valy, and Jean-Christophe Burie contributed to paper writing and editing.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## References


1. tranScriptorium. Available online: http://transcriptorium.eu/ (accessed on 20 February 2018).
2. READ Project—Recognition and Enrichment of Archival Documents. Available online: https://read.transkribus.eu/ (accessed on 20 February 2018).
3. IAM Historical Document Database (IAM-HistDB)—Computer Vision and Artificial Intelligence. Available online: http://www.fki.inf.unibe.ch/databases/iam-historical-document-database (accessed on 20 February 2018).
4. Ancient Lives: Archive. Available online: https://www.ancientlives.org/ (accessed on 20 February 2018).
5. Document Image Analysis—CVISION Technologies. Available online: http://www.cvisiontech.com/library/pdf/pdf-document/document-image-analysis.html (accessed on 20 February 2018).
6. Ramteke, R.J. Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition. *Int. J. Comput. Appl.* **2010**, *1*, 1–5. [CrossRef]
7. Siddharth, K.S.; Dhir, R.; Rani, R. Handwritten Gurmukhi Numeral Recognition using Different Feature Sets. *Int. J. Comput. Appl.* **2011**, *28*, 20–24. [CrossRef]
8. Sharma, D.; Jhajj, P. Recognition of Isolated Handwritten Characters in Gurmukhi Script. *Int. J. Comput. Appl.* **2010**, *4*, 9–17. [CrossRef]
9. Aggarwal, A.; Singh, K.; Singh, K. Use of Gradient Technique for Extracting Features from Handwritten Gurmukhi Characters and Numerals. *Procedia Comput. Sci.* **2015**, *46*, 1716–1723. [CrossRef]
10. Lehal, G.S.; Singh, C.A. Gurmukhi script recognition system. In Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, 3–7 September 2000; pp. 557–560.
11. Rothacker, L.; Fink, G.A.; Banerjee, P.; Bhattacharya, U.; Chaudhuri, B.B. Bag-of-features HMMs for segmentation-free Bangla word spotting. In Proceedings of the 4th International Workshop on Multilingual OCR, Washington, DC, USA, 24 August 2013; p. 5.
12. Ashlin Deepa, R.N.; Rao, R.R. Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review. *Int. J. Adv. Trends Comput. Sci. Eng.* **2014**, *3*, 481–485.
13. Kasturi, R.; O'Gorman, L.; Govindaraju, V. Document image analysis: A primer. *Sadhana* **2002**, *27*, 3–22. [CrossRef]
14. Paper History, Case Pap. Available online: http://www.casepaper.com/company/paper-history/ (accessed on 20 February 2018).
15. Doermann, D. *Handbook of Document Image Processing and Recognition*; Tombre, K., Ed.; Springer London: London, UK, 2014; p. 1055.
16. Chamchong, R.; Fung, C.C.; Wong, K.W. *Comparing Binarisation Techniques for the Processing of Ancient Manuscripts*; Nakatsu, R., Tosa, N., Naghdy, F., Wong, K.W., Codognet, P., Eds.; Springer Berlin: Berlin, Germany, 2010; pp. 55–64.
17. Kesiman, M.W.A.; Burie, J.-C.; Ogier, J.-M.; Wibawantara, G.N.M.A.; Sunarya, I.M.G. AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset. In Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016; pp. 168–172.
18. *The Unicode® Standard*; version 9.0—Core Specification; The Unicode Consortium: Mountain View, CA, USA, 2016.
19. Balinese Alphabet, Language and Pronunciation. Available online: http://www.omniglot.com/writing/balinese.htm (accessed on 20 February 2018).
20. Khmer Manuscript—Recherche. Available online: http://khmermanuscripts.efeo.fr/ (accessed on 20 February 2018).

21. Valy, D.; Verleysen, M.; Chhun, S.; Burie, J.-C. A New Khmer Palm Leaf Manuscript Dataset for Document Analysis and Recognition—SleukRith Set. In Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, 10–11 November 2017; pp. 1–6.
22. Suryani, M.; Paulus, E.; Hadi, S.; Darsa, U.A.; Burie, J.-C. The Handwritten Sundanese Palm Leaf Manuscript Dataset From 15th Century. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 796–800.
23. Kesiman, M.W.A.; Prum, S.; Burie, J.-C.; Ogier, J.-M. An Initial Study on the Construction of Ground Truth Binarized Images of Ancient Palm Leaf Manuscripts. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015; pp. 656–660.
24. Kesiman, M.W.A.; Prum, S.; Sunarya, I.M.G.; Burie, J.-C.; Ogier, J.-M. An Analysis of Ground Truth Binarized Image Variability of Palm Leaf Manuscripts. In Proceedings of the 5th International Conference Image Processing Theory Tools Application (IPTA 2015), Orleans, France, 10–13 November 2015; pp. 229–233.
25. Burie, J.-C.; Coustaty, M.; Hadi, S.; Kesiman, M.W.A.; Ogier, J.-M.; Paulus, E.; Sok, K.; Sunarya, I.M.G.; Valy, D. ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts. In Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016; pp. 596–601.
26. Kesiman, M.W.A.; Valy, D.; Burie, J.-C.; Paulus, E.; Sunarya, I.M.G.; Hadi, S.; Sok, K.H.; Ogier, J.-M. Southeast Asian palm leaf manuscript images: A review of handwritten text line segmentation methods and new challenges. *J. Electron. Imaging.* **2016**, *26*, 011011. [CrossRef]
27. Valy, D.; Verleysen, M.; Sok, K. Line Segmentation for Grayscale Text Images of Khmer Palm Leaf Manuscripts. In Proceedings of the 7th International Conference Image Processing Theory Tools Application (IPTA 2017), Montreal, QC, Canada, 28 November–1 December 2017.
28. Kesiman, M.W.A.; Prum, S.; Burie, J.-C.; Ogier, J.-M. Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images. In Proceedings of the 23rd International Conference Pattern Recognition, Cancun, Mexico, 4–8 December 2016; pp. 4017–4022.
29. Kesiman, M.W.A.; Burie, J.-C.; Ogier, J.-M. A Complete Scheme of Spatially Categorized Glyph Recognition for the Transliteration of Balinese Palm Leaf Manuscripts. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 125–130.
30. Bezerra, B.L.D. *Handwriting: Recognition, Development and Analysis*; Bezerra, B.L.D., Zanchettin, C., Toselli, A.H., Pirlo, G., Eds.; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2017; ISBN 978-1-53611-957-2.
31. Arica, N.; Yarman-Vural, F.T. Optical character recognition for cursive handwriting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 801–813. [CrossRef]
32. Blumenstein, M.; Verma, B.; Basli, H. A novel feature extraction technique for the recognition of segmented handwritten characters. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 3–6 August 2003; pp. 137–141.
33. O'Gorman, L.; Kasturi, R. *Executive briefing: Document Image Analysis*; IEEE Computer Society Press: Los Alamitos, CA, USA, 1997; p. 107.
34. Naveed Bin Rais, M.S.H. Adaptive thresholding technique for document image analysis. In Proceedings of the 8th International Multitopic Conference, Lahore, Pakistan, 24–26 December 2004; pp. 61–66.
35. Ntirogiannis, K.; Gatos, B.; Pratikakis, I. An Objective Evaluation Methodology for Document Image Binarization Techniques. In Proceedings of the Eighth IAPR International Workshop Document Annual System 2008, Nara, Japan, 16–19 September 2008; pp. 217–224.
36. He, J.; Do, Q.D.M.; Downton, A.C.; Kim, J.H. A comparison of binarization methods for historical archive documents. In Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, South Korea, 31 August–1 September 2005; pp. 538–542.
37. Gatos, B.; Ntirogiannis, K.; Pratikakis, I. DIBCO 2009: Document image binarization contest. *Int. J. Doc. Anal. Recognit.* **2011**, *14*, 35–44. [CrossRef]
38. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. ICDAR 2013 Document Image Binarization Contest (DIBCO 2013). In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1471–1476.
39. Howe, N.R. Document binarization with automatic parameter tuning. *Int. J. Doc. Anal. Recognit.* **2013**, *16*, 247–258. [CrossRef]

40. ICFHR2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts. Available online: http://amadi.univ-lr.fr/ICFHR2016_Contest/ (accessed on 20 February 2018).
41. Gupta, M.R.; Jacobson, N.P.; Garcia, E.K. OCR binarization and image pre-processing for searching historical documents. *Pattern Recognit.* **2007**, *40*, 389–397. [CrossRef]
42. Feng, M.-L.; Tan, Y.-P. Contrast adaptive binarization of low quality document images. *IEICE Electron. Express* **2004**, *1*, 501–506. [CrossRef]
43. Global Image Threshold Using Otsu's Method—MATLAB Graythresh—MathWorks France. Available online: https://fr.mathworks.com/help/images/ref/graythresh.html?requestedDomain=true (accessed on 20 February 2018).
44. Khurshid, K.; Siddiqi, I.; Faure, C.; Vincent, N. Comparison of Niblack Inspired Binarization Methods for Ancient Documents. In Proceedings of the Document Recognition and Retrieval XVI, 72470U, San Jose, CA, USA, 21 January 2009; p. 72470U. [CrossRef]
45. Sauvola, J.; Pietikäinen, M. Adaptive document image binarization. *Pattern Recognit.* **2000**, *33*, 225–236. [CrossRef]
46. Wolf, C.; Jolion, J.-M.; Chassaing, F. Text Localization, Enhancement and Binarization in Multimedia Documents. In Proceedings of the Object recognition supported by user interaction for service robots, Quebec City, QC, Canada, 11–15 August 2002; pp. 1037–1040.
47. Arvanitopoulos, N.; Susstrunk, S. Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts. In Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, Greece, 1–4 September 2014; pp. 726–731.
48. Hossain, M.Z.; Amin, M.A.; Yan, H. Rapid Feature Extraction for Optical Character Recognition. Available online: http://arxiv.org/abs/1206.0238 (accessed on 20 February 2018).
49. Fujisawa, Y.; Shi, M.; Wakabayashi, T.; Kimura, F. 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, Bangalore, India, 22 September 1999; pp. 277–280.
50. Kumar, S. Neighborhood Pixels Weights-A New Feature Extractor. *Int. J. Comput. Theory Eng.* **2009**, *2*, 69–77. [CrossRef]
51. Bokser, M. Omnidocument technologies. *Proc. IEEE.* **1992**, *80*, 1066–1078. [CrossRef]
52. Coates, A.; Lee, H.; Ng, A.Y. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 215–223.
53. Coates, A.; Carpenter, B.; Case, C.; Satheesh, S.; Suresh, B.; Wang, T.; Wu, D.J.; Ng, A.Y. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 440–445.
54. Shishtla, P.; Ganesh, V.S.; Subramaniam, S.; Varma, V. A language-independent transliteration schema using character aligned models at NEWS 2009. In Proceedings of the Association for Computational Linguistics, Suntec, Singapore, 7 August 2009; p. 40. [CrossRef]
55. Ul-Hasan, A.; Breuel, T.M. Can we build language-independent OCR using LSTM networks? In Proceedings of the 4th International Workshop on Multilingual OCR, Washington, DC, USA, 24 August 2013.
56. Ocropy: Python-Based Tools for Document Analysis and OCR, 2018. Available online: https://github.com/tmbdev/ocropy (accessed on 20 February 2018).
57. Homemade Manuscript OCR (1): OCRopy, Sacré Grl. Available online: https://graal.hypotheses.org/786 (accessed on 20 February 2018).
58. Breuel, T.M.; Ul-Hasan, A.; Al-Azawi, M.A.; Shafait, F. High-Performance OCR for Printed English and Fraktur Using LSTM Networks. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 683–687. [CrossRef]
59. Valy, D.; Verleysen, M.; Sok, K. Line Segmentation Approach for Ancient Palm Leaf Manuscripts using Competitive Learning Algorithm. In Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016.
60. Saund, E.; Lin, J.; Sarkar, P. PixLabeler: User Interface for Pixel-Level Labeling of Elements in Document Images. In Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 646–650. [CrossRef]

61. Stamatopoulos, N.; Gatos, B.; Louloudis, G.; Pal, U.; Alaei, A. ICDAR 2013 Handwriting Segmentation Contest. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1402–1406. [CrossRef]
62. PRImA. Available online: http://www.primaresearch.org/tools/Aletheia (accessed on 20 February 2018).
63. Clausner, C.; Pletschacher, S.; Antonacopoulos, A. Aletheia—An Advanced Document Layout and Text Ground-Truthing System for Production Environments. In Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 48–52. [CrossRef]