AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset

Made Windu Antara Kesiman, Jean-Christophe Burie, Jean-Marc Ogier, Gusti Ngurah Made Agus Wibawantara, I Made Gede Sunarya

AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset

Made Windu Antara Kesiman, Jean-Christophe Burie, Jean-Marc Ogier
Laboratoire Informatique Image Interaction (L3i)
University of La Rochelle, Avenue Michel Crépeau
17042, La Rochelle Cedex 1, France
{made_windu_antara.kesiman, jcburie, jean-marc.ogier}@univ-lr.fr

Gusti Ngurah Made Agus Wibawantara, I Made Gede Sunarya
Laboratory of Cultural Informatics (LCI)
Ganesha University of Education, Jalan Udayana No 11
81116, Singaraja, Bali, Indonesia
{agus.wibawantara, sunarya}@undiksha.ac.id

***Abstract*— We present the *AMADI_LontarSet*, the first handwritten Balinese palm leaf manuscript dataset. It includes three components of dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. The dataset was constructed from a hundred pages of randomly selected collections of palm leaf manuscripts from Bali, Indonesia. The dataset is publicly available for scientific use.**

Keywords-Balinese script; palm leaf manuscript; ground truth; image; dataset

I. INTRODUCTION

Ancient manuscripts record many important knowledges about world civilization histories. These manuscripts are a very valuable cultural heritage which contains a wide variety of social cultural life aspects. In Southeast Asia, most of the ancient manuscripts are discovered in the form of palm leaf manuscripts. They are written on a dried palm leaf by using a sharp pen (which looks like a small knife) and colored with natural dyes. The existence of palm leaf manuscripts in Southeast Asia is also very important both in term of quantity and variety of historical contents. It attracts the historians, philologists, and archeologists to discover more about the ancient ways of life. But unfortunately, there is only a limited access to the content of the manuscripts, because of the linguistic difficulties and the fragility of the document.

Therefore, the digitization and indexing projects for palm leaf manuscripts were proposed [1,2]. They work not only to digitize the palm leaf manuscripts, but also to develop an automatic analysis, transcription and indexing system for the manuscripts. The main objectives are to preserve the cultural heritages, and to open a wider access to the content of manuscripts for all scholars in the world. To achieve those objectives, ancient palm leaf manuscripts finally received great attention from researchers in the field of document image analysis [1ï 4]. Nowadays, the development of document analysis methods for palm leaf manuscripts is considered as a major challenge for handwritten document analysis. The challenges range wide from binarization process, character and text recognition tasks, to the word spotting methods.

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 was a necessary step for the research community. Under the scheme of the AMADI (Ancient Manuscripts Digitization and Indexation) Project, in this paper, we present the *AMADI_LontarSet*, the first handwritten Balinese palm leaf manuscript dataset. It includes three components of dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. The dataset is constructed from a hundred pages of randomly selected collections of palm leaf manuscripts from Bali, Indonesia.

This paper is organized as follow: Section II gives a brief description about Balinese script on the collection of palm leaf manuscripts and the challenges for document analysis. Section III presents a brief description about the corpus of palm leaf manuscripts and the digitization process. The ground truth construction for the *AMADI_LontarSet* is described in Section IV. Conclusions with some prospects for the future works are given in Section V.

II. PALM LEAF MANUSCRIPTS

A. The Collection of Palm Leaf Manuscripts from Bali, Indonesia

Bali has a great social and cultural history dating back several hundred years ago. Many literary texts of the Balinese were written on dried and treated palm leaves, called *Lontar*. The palm leaves are held and linked together by a string that passes through the central holes and knotted at the outer ends. *Lontars* store various forms of knowledge and historical record of the social life of Balinese cultures long ago. The content varies from ordinary texts to the most sacred writings. With a great influence from Indian culture, the Balinese manuscripts content were mostly based on the famous Indian epics of Ramayana and Mahabharata. Many

Lontars contain information on important issues such as medicines and village regulations that are used as daily guidance, including texts on religion, holy formulae, rituals, family genealogies, law codes, treaties on medicine, arts and architecture, calendars, prose, poems and even magics. But unfortunately, many *Lontars* discovered are the collection of the museum and private family that has been in a state of disrepair due to age and due to inadequate storage conditions.

B. The Writing in Balinese Script

To create a *Lontar*, the texts were inscribed with a small knife-like pen (a special tool called *Pengerupak*). It is made of iron, with its tip sharpened in a triangular shape so it can make both thick and thin inscriptions. The manuscripts were then scrubbed by a natural dyes to leave a black color on the scratched part as text. The Balinese palm leaf manuscripts were written in Balinese script in Balinese language. Writing in Balinese script, there is no space between words in a text line. Some characters are written on upper baseline or under the baseline of text line (Fig. 1). *Lontars* were written in the ancient literary texts composed in the old Javanese language of Kawi and Sanskrit. Balinese script is considered to be one of the complex scripts from Southeast Asia. The alphabet and numeral of Balinese script is composed of ±100 character classes including consonants, vowels, diacritics, and some other special compound characters. In reality, the majority of Balinese have never read any *Lontar* because of language obstacles as well as tradition which perceived them as sacrilege.

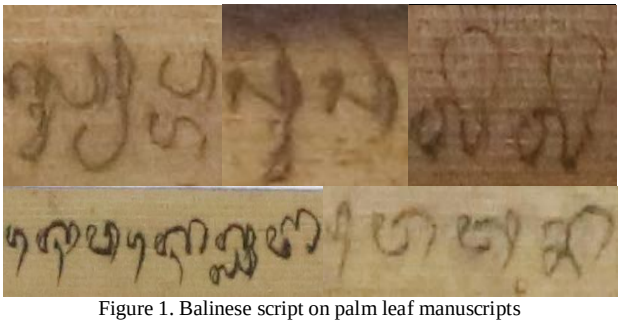

Figure 1. Balinese script on palm leaf manuscripts

C. The Challenges of Document Analysis for Palm Leaf Manuscript Images

Due to its specific characteristics, palm leaf manuscripts are providing new challenges in document analysis. Usually, palm leaf manuscripts are of poor quality since the documents have degraded over time due to storage conditions (Fig. 2). Natural materials from palm leaves certainly cannot ýght against time, and therefore the process of digitizing and indexing *Lontars* are very important. The palm leaf manuscripts contain discolored parts and artefacts due to aging and low intensity variations or poor contrast, random noises, and fading [1]. Several deformations in the character shapes are visible due to the merges and fractures of the use of nonstandard fonts, varying space between letters, and varying space between lines. It is known that the similarities of distinct character shapes, the overlaps, and interconnection of the neighboring characters further complicate the problem of OCR system [5]. One of the main problem faced when dealing with segmented handwritten character recognition is the ambiguity and illegibility of the characters [6]. This characteristics provide a suitable challenge for testing and evaluation of robustness for feature extraction methods which were already proposed for character recognition. Balinese scripts on palm leaf manuscripts offer a real new challenge in document analysis system development.

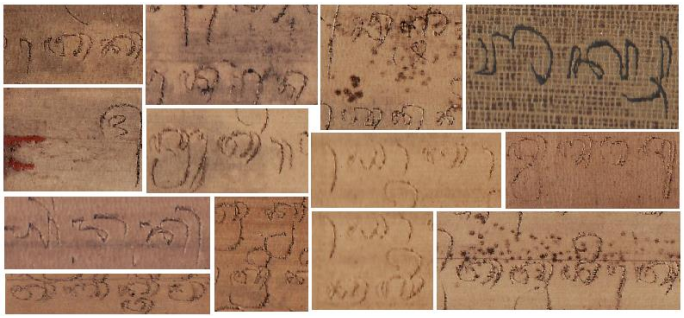

Figure 2. The degradations on palm leaf manuscripts [1]

III. CORPUS AND DIGITIZATION PROCESS

A. Location and Source of Collection

Our first corpus of palm leaf manuscript images are the sample images of the palm leaf manuscripts from Bali, Indonesia. It is hard to estimate the number of the whole collection of palm leaf manuscript in Bali because most of the palm leaf manuscript collections are kept by the private family 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 come from 5 different locations (regions): 2 museums and 3 private families. It consists of randomly selected 10 collections from Museum Gedong Kertya, 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 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 I.

B. Camera and Digitization Support

To capture the manuscript images, we used a Canon EOS 5D Mark III camera. The camera settings are as follows [2]: F-stop: f/22 (diafragma), 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 *óOnô*. We also designed a black box camera support by wood to avoid the irregular lighting/luminance condition and to fits our semi outdoor capturing location (Fig. 3). This camera support was optimally designed to be used under some restricted conditions given by the museum or the owner of the manuscripts. Two additional light are added inside the black box support with White Neon 50 cm 20 watt.

Thumbnail samples of the captured images are showed in Fig. 4. To digitize large collections of palm leaf manuscript and to place them online, the philologists consider the quality of these images are good enough.

TABLE I. COLLECTION OF PALM LEAF MANUSCRIPTS FROM BALI, INDONESIA

Location	Collection Code	Content	Nb of captured pages
Museum Gedong Kertya, 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, 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, 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, Bangli (1 collection)	Bangli	*Sabung Ayam*	82
Village of Rendang, Karangasem (1 collection)	WN	*Surat Jual Beli Tanah*	24
TOTAL			393

IV. GROUND TRUTH CONSTRUCTION

A. *Binarized Images Ground Truth Dataset*

Binarization process is one of the early and important stage in document analysis pipeline, is also a real challenge for palm leaf manuscripts. Some document analysis methods still require a good binarized image as a preliminary condition. In order to evaluate and to select an optimal binarization method, the ground truth binarized image is needed. Therefore, creating the binarized images ground truth dataset for palm leaf manuscripts is a necessary step in our research.

In our previous work [1], we proposed a specific semi-local binarization scheme to overcome the ground truth creation difficulty on degraded and low quality palm leaf manuscript images (Fig. 5). This scheme will help as the initial binarization process for the semi-automatics framework for construction of ground truth binarized image (Fig. 6). This framework is based on the one used to build the database in DIBCO competition series [7].

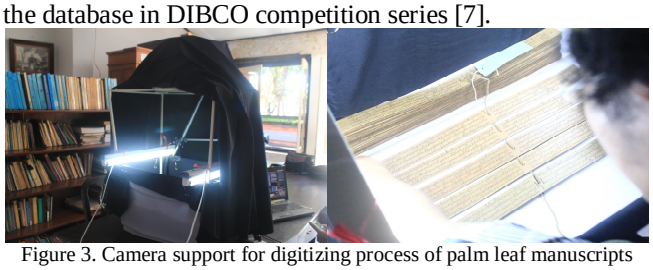

Figure 3. Camera support for digitizing process of palm leaf manuscripts

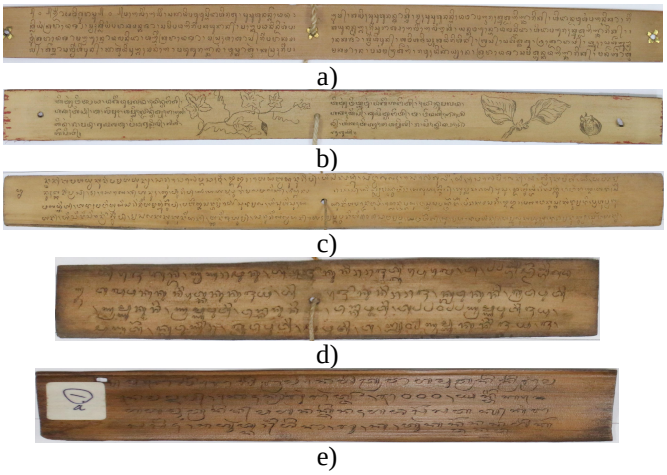

Figure 4. Sample images of palm leaf manuscript from a) Museum Gedong Kertya, Singaraja, b) Museum Bali, Denpasar, c) Village of Jagaraga, Buleleng, d) Village of Susut, Bangli, e) Village of Rendang, Karangasem

The idea of our ó*semi-local*ô concept is to apply a powerful global binarization method on only precise local character area. The first initial binarization process is needed to optimally separate text from the background, and it provides a first binary image for the skeletonizing process of the characters on manuscript.

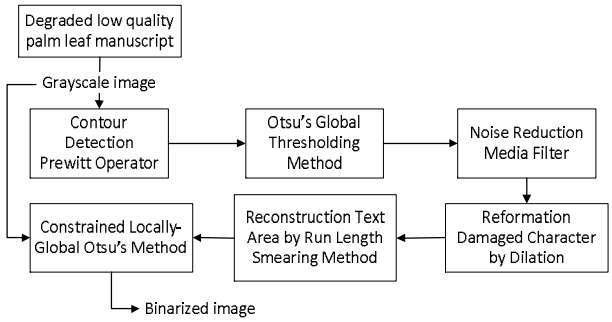

Figure 5. Semi-local binarization scheme [1]

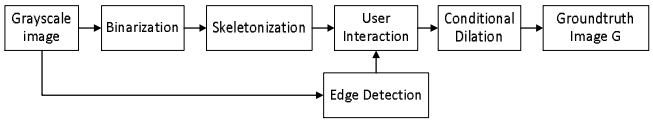

Figure 6. Ground truth construction procedure used for DIBCO series [7]

In this framework, human intervention plays a very important task by performing the manual correction of the character skeletons based on character edges (detected using Canny algorithm [8]). In order to purely measure the variability of human subjectivity in our ground truth creation, for some images, we did not apply any initial binarization and skeletonization methods. The skeletonization process is completely performed by human. More than 70 students from the Department of Informatics Education, Ganesha University of Education, Singaraja Bali, were asked to trace manually the skeleton of the Balinese character found in palm leaf manuscript image with PixLabeler tool [9]. One student worked with two different images, and one image was ground truthed by two different students. These two manually skeletonized image will be re-skeletonized with Matlab function *bwmorph*[1] to make sure that the skeleton is only one pixel wide.

The final estimated ground truth binarized image is then automatically constructed by dilating the skeleton image, constrained by the character edges. The skeleton is dilated until Canny edges intersect each binarized component of the dilated skeleton in a ratio of 0.1. This value of minimal ratio between number of pixels in intersection of Canny edge and number of pixels of the dilated skeleton is found based on our empirical experiment and observation on the thickness of the character stroke in our manuscripts [2]. Table II shows the summary of binarized images ground truth dataset for the *AMADI_LontarSet*. For the training-based binarization method, we divide our dataset into two subset: 50 images for training and 50 images for testing. Fig. 7 shows some samples of binarized images ground truth from our dataset. For more detail about the analysis of ground truth binarized image variability of palm leaf manuscripts, please refer to our previous work in [2].

TABLE II. SUMMARY OF BINARIZED IMAGES GROUND TRUTH DATASET FOR THE AMADI_LONTARSET

No.	Data	Format	Qty.
1.	Original Images of Manuscript	RGB Color image - JPG	100 images
2.	Binarized Ground Truth Image (1st ground truther)	Binary image - BMP	100 images
3.	Binarized Ground Truth Image (2nd ground truther)	Binary image - BMP	100 images

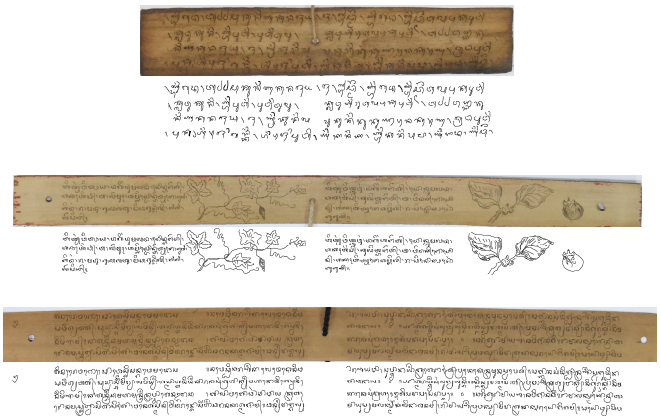

Figure 7. Samples of binarized images ground truth dataset

B. *Word Annotated Images Dataset*

To create the word annotated ground truth dataset of the manuscript, we asked a collaborative work between the Balinese script philologists, students from the Department of Informatics, and students from the Department of Balinese Literature. The philologists read the manuscripts and create the Latin transcription. Based on this Latin transcription, a pair of students (one student in Informatics and one student in Balinese Literature) works together to segment and to annotate each word in manuscripts. The validation and correction of word annotation are done based on the expertise of the philologists. Any further discussion remains open between the philologists and the ground truthers to correct and to validate the transcription while the annotation process.

We used ALETHEIA[2], an advanced document layout and text ground-truthing system [10], to segment and to annotate the words (Fig. 8). After the segmentation and the annotation process, the manuscript images are then cropped based on word polygon coordinates in the XML file produced by ALETHEIA (Fig. 9). For all word annotated images in this dataset, we use the filename format as follows:

***word_filename_idword_cTL_rTL_cBR_rBR**.jpg*

where ***word*** indicates the word string of this word segment, ***filename*** indicates the original manuscript image of this word segment, ***idword*** indicates the id of this word segment (used only for Aletheia software), ***cTL*** indicates column coordinates of top left point for this word segment, ***rTL*** indicates row coordinates of top left point for this word segment, ***cBR*** indicates column coordinates of bottom right point for this word segment, ***rBR*** indicates row coordinates of bottom right point for this word segment. The image coordinates for column=1 and row=1 are considered as the pixel on the top left corner of the image. Table III shows the summary of word annotated images dataset for the *AMADI_LontarSet*.

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

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

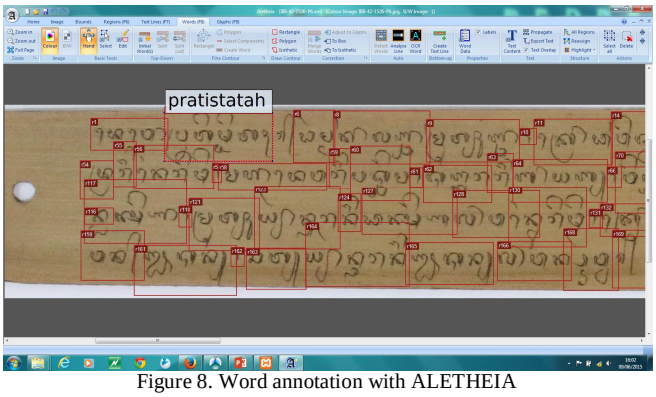

Figure 8. Word annotation with ALETHEIA


Figure 9. Samples of word annotated images

TABLE III. SUMMARY OF WORD ANNOTATED IMAGES DATASET FOR THE AMADI_LONTARSET

No.	Data	Format	Qty.
1.	Training Set: Original Images of Manuscript	RGB Color image - JPG	130 images
2.	Training Set: Transcription of manuscript of No 1	TXT	130 filetexts
3.	Training Set: Word annotated images of No 1	RGB Color image - JPG	15,022 images
4.	Testing Set: Original Images of Manuscript	RGB Color image - JPG	100 images
5.	Testing Set: Transcription of manuscript of No 4	TXT	100 filetexts
6.	Testing Set: Word annotated images of No 4	RGB Color image - JPG	10,475 images
7.	Testing Set: Selected word annotated images as query-by-example	RGB Color image - JPG	36 images
8.	Testing Set: Ground truth images for all query images of No 7	RGB Color image - JPG	257 images

C. *Isolated Character Annotated Images Dataset*

By using the collection of word annotated images which were produced in our previous ground truthing process, we collected our isolated handwritten Balinese character dataset. First, we applied Otsu [11,12] binarization method to all word patch images. We automatically extracted all connected component found on 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 (Fig. 10). This interface shows all character segments which are automatically segmented from each word annotated image based on character component extraction. 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 are also displayed in group based on their given class. A hyperlink from each annotated character to their corresponding word annotated images is provided to allow the philologists to verify and to correct the annotation (Fig. 11).

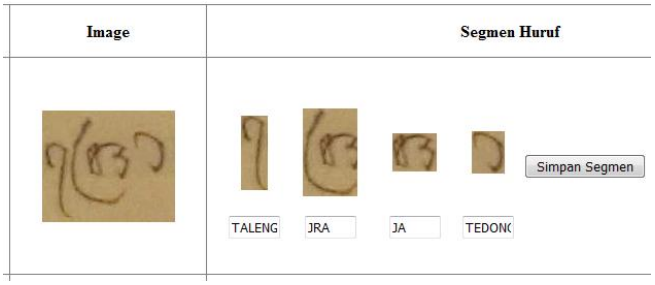

Figure 10. Screenshot of web based user interface for the character annotation process

Figure 11. Screenshot of character class verification

All patch images that have been segmented and annotated at character level will serve as our isolated character dataset. Table IV shows the summary of isolated character annotated images dataset for the *AMADI_LontarSet*. The number of sample images for each classes is different. Some classes are frequently found in our collection of palm leaf manuscripts, but some others are rarely used. Thumbnail samples of these character annotated images are showed in Fig. 12.

TABLE IV. SUMMARY OF BINARIZED IMAGES GROUND TRUTH DATASET FOR THE AMADI_LONTARSET

No.	Data	Format	Qty.
1.	Training Set: Character annotated images	RGB Color image - JPG	133 classes 11,710 images
2.	Testing Set: Character annotated images	RGB Color image - JPG	133 classes 7,673 images

Figure 12. Samples of character-level annotated patch images of Balinese script on palm leaf manuscripts

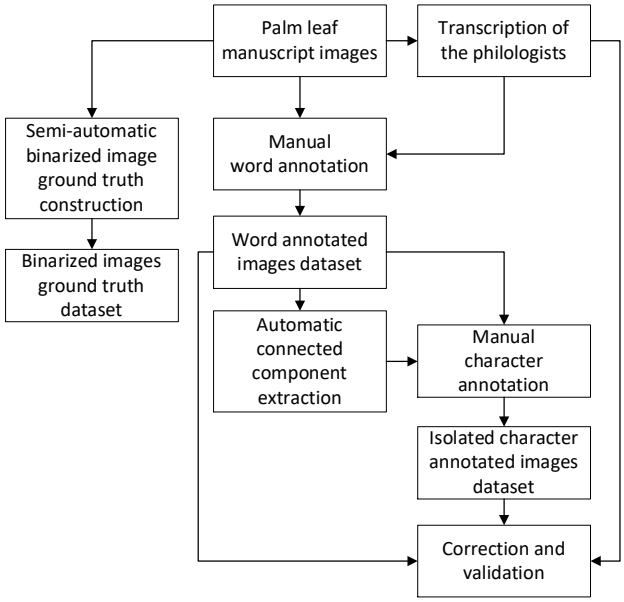

Figure 13. Overall scheme of ground truth dataset construction

Fig. 13 shows the overall scheme of ground truth construction for all dataset of palm leaf manuscript images.

V. Conclusions and Future Works

The development of document analysis methods for palm leaf manuscripts is considered as a major challenge for handwritten document analysis. In order to develop and to evaluate the performance of the document analysis methods, the dataset and the corresponding ground truth data for palm leaf manuscripts is necessary. We present the *AMADI_LontarSet*, the first handwritten Balinese palm leaf manuscript dataset. It includes three components of dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. The whole dataset will be publicly available for scientific use after the ICFHR 2016 conference on http://amadi.univ-lr.fr/ICFHR2016_Contest/. For the future works, we will develop the dataset in term of data quantity and variety to be able to provide sufficiently a larger train data set for document analysis methods.

Acknowledgment

The authors would like to thank Museum Gedong Kertya, Museum Bali, and all families in Bali, Indonesia, for providing us the samples of palm leaf manuscripts, and the students from the Department of Informatics Education and the Department of Balinese Literature, Ganesha University of Education for helping us in ground truthing process for this research project. This work is also 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).

References

[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.

[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. 229ï 233.

[3] R. Chamchong, C.C. Fung, Character segmentation from ancient palm leaf manuscripts in Thailand, in: ACM Press, 2011: p. 140 doi:10.1145/2037342.2037366.

[4] R. Chamchong, C.C. Fung, K.W. Wong, Comparing Binarisation Techniques for the Processing of Ancient Manuscripts, in: R. Nakatsu, N. Tosa, F. Naghdy, K.W. Wong, P. Codognet (Eds.), Cult. Comput., Springer Berlin Heidelberg, Berlin, Heidelberg, 2010: pp. 55ï 64. http://link.springer.com/10.1007/978-3-642-15214-6_6 (accessed December 5, 2014).

[5] N. Arica, F.T. Yarman-Vural, Optical character recognition for cursive handwriting, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 801ï 813. doi:10.1109/TPAMI.2002.1008386.

[6] M. Blumenstein, B. Verma, H. Basli, A novel feature extraction technique for the recognition of segmented handwritten characters, in: IEEE Comput. Soc, 2003: pp. 137ï 141. doi:10.1109/ICDAR.2003.1227647.

[7] K. Ntirogiannis, B. Gatos, I. Pratikakis, Performance Evaluation Methodology for Historical Document Image Binarization, IEEE Trans. Image Process. 22 (2013) 595ï 609. doi:10.1109/TIP.2012.2219550.

[8] J. Canny, A Computational Approach to Edge Detection, IEEE Tranaction on Pattern Analysis and Maching. Intelligence. PAMI-8 (1986) 679ï 698. doi:10.1109/TPAMI.1986.4767851.

[9] E. Saund, J. Lin, P. Sarkar, PixLabeler: User Interface for Pixel-Level Labeling of Elements in Document Images, in: IEEE, 2009: pp. 646ï 650. doi:10.1109/ICDAR.2009.250.

[10] C. Clausner, S. Pletschacher, A. Antonacopoulos, Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments, in: IEEE, 2011: pp. 48ï 52. doi:10.1109/ICDAR.2011.19.

[11] I. Pratikakis, B. Gatos, K. Ntirogiannis, ICDAR 2013 Document Image Binarization Contest (DIBCO 2013), in: IEEE, 2013: pp. 1471ï 1476. doi:10.1109/ICDAR.2013.219.

[12] I.B. Messaoud, H. El Abed, V. Märgner, H. Amiri, A design of a preprocessing framework for large database of historical documents, in: ACM Press, 2011: p. 177. doi:10.1145/2037342.2037372.