A New Khmer Palm Leaf Manuscript Dataset for Document Analysis and Recognition – SleukRith Set

Dona Valy[a,b], Michel Verleysen[a], Sophea Chhun[b], and Jean-Christophe Burie[c]
[a]ICTEAM Institute, Université Catholique de Louvain, Belgium
[b]Department of Information and Communication Engineering, Institute of Technology of Cambodia, Cambodia
[c]Laboratoire Informatique Image Interaction (L3i), University of La Rochelle, France
{dona.valy,michel.verleysen}@uclouvain.be, sophea.chhun@itc.edu.kh, jcburie@univ-lr.fr

ABSTRACT

Analysis of ancient Khmer documents can be quite challenging due to the elaborated shape of Khmer handwritten characters combined with the complex structure of how words are formed from those characters. Palm leaf manuscripts, one of the most well-known old Khmer documents, have been being digitized and centralized; therefore, document analysis functions such as text search capabilities are necessary but still remain unavailable for this type of documents. In order to contribute to the progress of relevant researches, we introduce in this paper a new dataset called SleukRith set comprising of 657 pages of Khmer palm leaf manuscripts randomly selected from various collections whose quality and digitization method are variable. The dataset contains three types of data: isolated characters, words, and lines. Each type of data is annotated with the ground truth information which is very useful for evaluating and serving as a training set for common document analysis tasks such as character/text recognition, word/line segmentation, and word spotting. In order to serve as a base line, the result of an evaluation study of Khmer isolated character recognition that we have conducted on SleukRith Set using Convolutional Neural Network is also presented.

CCS CONCEPTS

• Information systems~Digital libraries and archives • Applied computing~Optical character recognition

KEYWORDS

Handwritten document analysis, ground truth, annotated dataset; palm leaf manuscript

HIP2017, November 10–11, 2017, Kyoto, Japan

ACM ISBN 978-1-4503-5390-8/17/11$15.00
https://doi.org/10.1145/3151509.3151510

1 INTRODUCTION

Before the invention of paper, palm leaves were one of the earliest form of writing mediums in numerous parts of Asia especially in many Southeast Asian countries. Nowadays in those countries, palm leaf manuscripts are considered to be of important cultural value. The contents of these ancient manuscripts are related to many fields of study such as history, religion, art, and even scientific discovery including mathematics, architecture, and medicine; therefore, they can still be very useful as a reference source for many researchers.

Due to aging, palm leaf manuscripts and consequently the knowledge contained in them are facing destruction. In the last few decades, many preservation programs for palm leaf documents have emerged. Besides preserving the manuscripts in their physical form, digital imaging through scanning and photography has been used to protect, centralize, and provide easier access of the documents to public. For instance, several online databases of digital palm leaf manuscript images such as the digital libraries from Lao [1], Northern Thailand [2], and Cambodia [3] have been made available. Nevertheless, only bibliometric indexing can be offered beside the image files: identification of the manuscript origin, title, date, topic, etc., but no indexing of the content is available yet. In-text search by keywords for example is still impossible. An automatic analysis, a transcription process, and an indexing system are therefore very much needed.

In recent years, research in document analysis on ancient palm leaf manuscripts has received considerable attention, for instance, the research work on layout analysis of palm leaf scripts including character/line segmentation [1-4] and recognition of isolated handwritten characters [5]. To develop, and evaluate new methods, a standard dataset is mandatory. An example of such dataset is the AMADI_LontarSet [6] which has been created with handwritten Balinese palm leaf manuscripts from Indonesia. Also, it is shown that text image datasets differ greatly by the language and the type of script, and to the best of our knowledge, no dataset comprising of annotated data for ancient Khmer handwritten scripts has been publicly made available yet.

[1] http://laomanuscripts.net
[2] http://lannamanuscripts.net
[3] http://khmermanuscripts.efeo.fr

In this paper, SleukRith Set, the first dataset specifically created for Khmer palm leaf manuscripts, is introduced. The dataset consists of annotated data from 657 pages of digitized palm leaf manuscripts which are selected arbitrarily from a large collection of existing and also recently digitized images.

The remainder of this paper is organized as follows. Section II gives a brief description of important characteristics of Khmer palm leaf documents and new challenges they create for the field of document analysis. Section III presents how a digital corpus of Khmer palm leaf manuscripts is built. An overall overview of SleukRith Set including the construction of its ground truth is described in details in Section IV. Section V covers an evaluation study of isolated character recognition on the dataset. Finally, a conclusion is given in Section VI.

2 CHARACTERISTICS OF KHMER PALM LEAF MANUSCRIPTS

2.1 Ancient Khmer Palm Leaf Scripts

Palm leaf writing has been passed on traditionally from generations to generations through scholars and scribers. When an old palm leaf document ages and decays, it is a common practice that its content is transferred to fresh new leaves. Khmer palm leaf manuscripts are written on long rectangular cut and dried palm leaf sheets. A metal stylus with sharp and pointed edge is used to make an incision in the shape of handwritten characters one at a time before black ink, which is a mixture of coal and a type of paste, is applied to make the carving noticeable and easy to read. Each palm leaf sheet typically has one hole in the middle or two holes at a distance of around one third of the length of the page from the left and the right ends. The sheets are tied together by passing a string through those holes to create a binding like a book.

The writing in most of Khmer palm leaf manuscripts uses Khmer alphabet which differs slightly according to the era during which the documents were created. The alphabet is composed of more or less 70 symbols including consonants, different types of vowels, diacritics, and special characters.

The languages written on the palm leaf documents vary from Khmer, the official language Cambodian people speak nowadays with slightly different spelling vocabularies, to Pali and Sanskrit by which the modern Khmer language is considerably influenced. Only a minority of Cambodian people such as philologists and Buddhist monks are able to read and understand the latter languages.

2.2 New Challenges for Document Analysis

Despite the availability of advanced image capturing methods, photography technology, and scanning equipment, the quality of many palm leaf images is still low due to natural aging and deterioration. Common degradations and noises include seepage of ink or bleed through, damage or tear around the area of the holes used for binding the document, stain from dirt, and other types of discoloration. Fig. 1 illustrates these degradations. The fragility of the aging leaves also presents some difficulties during the digitization process of the leaves. For instance, sometimes leaf manuscripts are curved and cannot be forced flat which results in an uneven illumination in the output image.

The characteristic of Khmer alphabet, which consists of a large number of symbols is also a big challenge attributable to the irregular position of how those characters are combined into words. Unlike the writing of most Latin languages where characters are sequenced from left to right, in Khmer writing, vowels are positioned either on the left, on the right, below, or above the consonant they are spelled with. Two or more consonants can also be merged together by transforming into different shapes called low-consonant or subscript form which are placed below the main consonant. These variations of how Khmer letters are formed produce multiple levels (sometimes more than three) of characters. Moreover, some writers tend to exaggerate their writing by elongating the upper or lower part of a character which makes it go far out of its main line, touch, or overlap with other characters from adjacent lines.

In addition, the ambiguity of certain characters in Khmer alphabet is a big challenge as well for character recognition problem. Some groups of characters can only be distinguishable by a mere difference of a small hole or a short stroke. Some types of symbols contain multiple parts whose shapes are identical or very similar to other characters. Fig. 2 shows some examples of this ambiguity.

In Khmer writing, even the modern one, there is no word separation. Spaces are occasionally used to separate phrases instead of words. We rely on grammatical structures, character combination rules, and sometimes contextual meaning of the sentence in order to identify where the beginning and the ending of a word are.

Text lines in palm leaf manuscripts may also be slanted or curved upward/downward on account of it being handwritten or from improper digitizing. Furthermore, holes used for binding the manuscript leaves leave behind empty areas in the middle of the page creating discontinuity of text lines.

Figure 1: Several types of deformations and defects found in palm leaf manuscripts

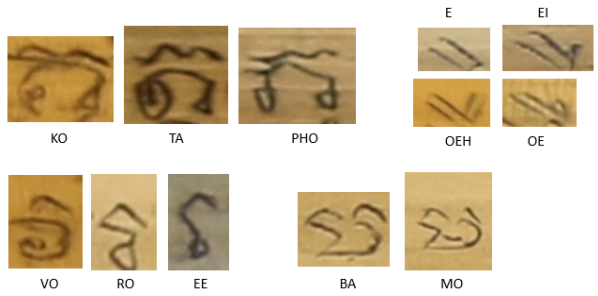

Figure 2: Example of similarity between certain Khmer characters

3 DIGITAL IMAGE CORPUS

3.1 Digitized Images from Various Establishments

In Cambodia, existing digital data of palm leaf manuscripts can be found in various libraries and institutions. Table I shows the number of digital image collections (a collection here refers to one complete set of palm leaf document) available in those establishments. Sample images are shown in Fig. 3.

3.1.1 École Française d'Extrême-Orient. The online database of Khmer manuscripts is the result of the work conducted by the École Française d'Extrême-Orient (EFEO-FEMC) research team since 1990, aiming to provide a comprehensive inventory and photographic collection of Cambodia's manuscripts [7]. The website is accessible to public and is home to hundreds of collections of palm leaf manuscripts. The digital images were captured from microfilms hence their low quality.

3.1.2 Buddhist Institute. The Buddhist Institute was an initiative of King Sisovat in 1921, when he inaugurated the royal library, Khemra Bannalai which subsequently changed its name to Preah Raj Bannalai in 1925 [8]. Then in 1930, King Monivong established the Buddhist Institute. The responsibilities of the Institute are not only research on Cambodian literature, language and Buddhism, but also publication and education. Only one collection containing 96 pages is available from this institute. The digitization method of this collection is unknown.

3.1.3 Phnom Penh National Library. The National Library of Cambodia was inaugurated by the French colonial administration [9]. Thereafter it was successively managed by French staff until the appointment of the first Khmer Director in 1951. After independence in 1954 there was a steady growth in Cambodian publishing, which was reflected in the increased number of Khmer language books in the National Library. Closed down during the Khmer Rouge era, the National Library was used for several years as accommodation by members of the Pol Pot regime, who destroyed many of the books. Since 1980 the National Library has been re-established with the assistance of various overseas governments and agencies. Today the National Library of Cambodia holds hundred thousands of copies in various languages (Khmer, French, English, and German). There is also a large collection of palm leaf manuscripts. Some collections of the manuscripts were digitized recently by a Khmer manuscript conservation and research group. We obtained in total 35 collections of already digitized manuscripts from this establishment.

Table 1: Number of Digital Image Collections Available in Various Establishments

Nº	Establishment	Digitization Method	Nb. Of Collections
1	École Française d'Extrême-Orient (EFEO)	Nikon F3	937
2	Buddhist Institute (BI)	Unknown	1
3	National Library (NL)	Canon 750D	35

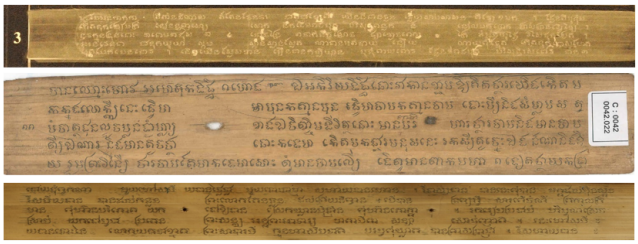

Figure 3: Samples of digitized images from top to bottom: EFEO, BI, and NL

3.2 Our Digitization Campaign

We also conduct our own digitization campaign in order to capture and collect palm leaf manuscript images found in Buddhist temples in different locations throughout Cambodia. A standard digitization procedure and a proper set up have been developed. Due to the fact that palm leaf manuscripts are fragile, and certain scripts are not allowed to be moved from the place where they are stored, digitizing using a scanner is not viable, so a portable option is needed. In order to capture the scripts, a Canon EOS 5DS professional camera is used. The camera settings are as follows: F-stop: f/4, shutter speed: 1/10 second, ISO: 100, focal length: 45 mm, distance to object: 65 cm, and auto focus: on. To support the camera to be able to shoot downward, we use a Manfrotto 055XPro3 tripod with its fluid head. To avoid irregular lighting condition and to adapt to our semi indoor/outdoor capturing location, the camera is covered over by a black cloth. Additional rechargeable led lights (led 715) are attached to each of the tripod legs.

Our campaign has been conducted in three locations in Cambodia: Phnom Penh, Kandal, and Siem Reap. We have collected and digitized manuscripts found mostly in Buddhist temples (pagodas). A summary of the collection from our digitization campaign is listed in Table II. Some sample images are shown in Fig. 4.

4 OVERVIEW OF SLEUKRITH SET

SleukRith set is a collection of three types of annotated data: isolated characters, words, and lines. The annotation is made on 657 pages of Khmer palm leaf manuscript randomly selected

from different sources. A summary of selected pages and their sources are shown in Table III. The majority of the images are chosen from the recently digitized manuscripts at the National Library and from our digitization campaign. Due to their low quality, the dataset only contains five pages each from the collections of EFEO and the Buddhist Institute.

Table 2: Collection of Digitized Palm Leaf Manuscripts from Our Digitization Campaign

Nº	Location	Nb. Of Collections	Nb. Of Pages
1	Tuol Tom Poung, Phnom Penh	2	54
2	Tek Vil Pagoda, Kandal	2	98
3	Bo Pagoda, Siem Reap	9	59
	Total	**13**	**211**

Figure 4: Sample images of our digitization campaign (from top to bottom: Phnom Penh, Kandal, and Siem Reap)

For annotating the three types of data, a tool with an easy-to-use user interface has been developed. The tool is implemented in Java hence its multi-platform portability. Since annotating a large amount of data can be quite exhausting, most interactions between the user and the tool are performed using only left and right clicks of mouse buttons. A keyboard input is required when giving labels to the data, modifying data, or deleting data.

The annotation process was accomplished with the help from 34 volunteer students from the department of Computer Science at the Institute of Technology of Cambodia (ITC) and the National Institute of Posts, Telecommunications, and ICT (NIPTICT). Each participant played the role of a ground truther and was assigned a set of palm leaf document images identified by number codes. Using their common knowledge of Khmer language, the ground truthers were asked to annotate each of their assigned pages according to the following steps: segment and label all characters, group segmented characters into words, and finally assigned each segmented character to a line it belongs to. After the initial annotation stage done by the students, a final validation and correction iteration has been performed verifying that the data is consistent and without errors.

Table 3: Collection of Palm Leaf Manuscript Images from Different Sources Composing SleukRith Set

Nº	Source	Nb. Of Pages
1	National Library	427
2	EFEO	26
3	Buddhist Institute	15
4	Our Digitization Campaign	189
	Total	**657**

4.1 Labeled Individual Character Dataset

Individual or isolated character dataset is the most important data type in SleukRith Set since its information is used to produce the other types of data. In order to segment and annotate a manuscript page into small image patches representing each individual character, a polygon boundary enclosing the character needs to be drawn manually. The ground truther is required to dot out vertex of the polygon one by one until a proper boundary is formed (see Fig. 5). The ground truther is then prompted to input the correct Unicode or Unicode sequence as label for that character.

A problem in annotating a character occurs when it is composed of multiple parts. In this situation, each part of the character is segmented separately and labeled with the original Unicode of the character followed by a number representing that part. A different situation is when multiple characters are merged together and form a new shape. In this case, the shape is then annotated as a whole and is given the label which is a sequence of the Unicode of the characters comprising it. Examples illustrating these cases are shown in Fig. 6.

Certain writers exaggerate their writing by elongating the ending stroke of the characters. Also some characters are written in a way that they encircle other characters. When cropped into a rectangular area, the image patch of such elaborate character does not only contain the character itself but also parts or the entirety of other characters. To solve this issue, by using the polygon boundary as a mask, an inpainting technique [10] can be applied in order to eliminate the unwanted area in the image patch. In Fig. 7, the image patch of character SUBYO is inpainted resulting in a new clean image.

4.2 Annotated Word Dataset

After all characters in the page are manually annotated, they can be combined together into words. To form a word, the character components of that word are selected one by one. The selection order is also important since Khmer Unicode sequence does not follow the left to right position order of the characters but instead respects a consonant-first-vowel-second basis. Fig. 8 shows an example illustrating this phenomenon. In this example, the word is composed of five characters, and the correct sequence is PHO-SUBLO-EE-CHA-BANTAK. Even though the vowel EE is at the left-most position of the word, it is placed third in the Unicode sequence after the consonant PHO and the sub-consonant SUBLO.

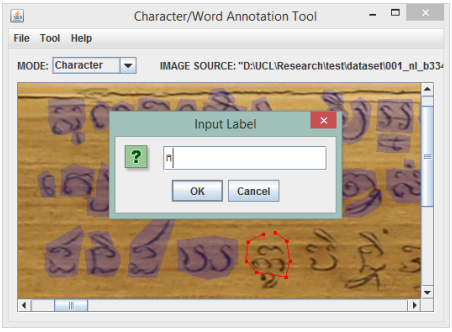

Figure 5: Annotation of individual character dataset

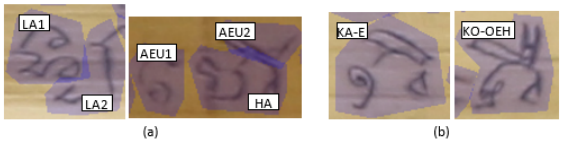

Figure 6: Examples of (a) characters containing multiple parts and (b) merged shapes

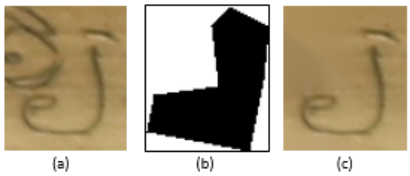

Figure 7: Application of inpainting technique on a character image patch (a) input image, (b) inpainting mask using polygon boundary, (c) result

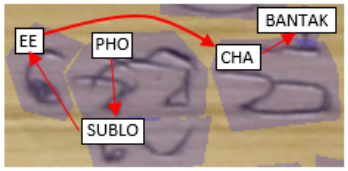

Figure 8: Order sequence of characters in a word

The ground truther is then again prompted to input a Unicode sequence representing the label of the formed word. By default, the word label is generated by putting together the labels of the characters which are the components of that word. The second label should also be provided by the ground truther when either the current word spelling is found to be erroneous or when an equivalent word from the modern Khmer language has a different spelling.

4.3 Line Segmentation Ground truth Dataset

Similarly annotated characters may be grouped into lines. To efficiently achieve this, the ground truther uses left click and drag over characters belonging to the same line. He is then asked to create a new line from the selected characters or add them to existing lines.

After all steps in the annotation scheme are complete, an xml file containing all information of the three types of data of the annotation can be exported for each manuscript page. It also serves as a temporary save file to store incomplete progress of the annotation. The xml file is divided into two sections (see Fig. 9). The upper part under the tag name "CharAnno" is dedicated to the annotation at the character level. This section block contains child blocks. Each child block represents an annotated character, information about the coordinates of its polygon boundary and additional attributes including character id, its label, and the id of the line which the character belongs to. The lower part of the file under the tag name "WordAnno" describes the annotation at the word level. Since a word is a combination of characters, only the id's of the annotated characters defined in the first section are stored along with the id information of the annotated word and its two labels. From this xml file, image patches representing characters and words can be generated (see Fig. 10 and Fig. 11). Table IV shows the current statistics of SleukRith Set.

5 EVALUATION OF ISOLATED CHARACTER RECOGNITION

In order to justify how the dataset can be used and to serve as a baseline, an evaluation study of isolated character recognition has been carried out on the individual character dataset of SleukRith Set. Small characters such as punctuation and diacritics are removed from the set resulting in 111 classes divided into training (113,206 samples) and test (90,669 samples) sets.

A multilayer convolutional neural network (CNN) is used in this experiment (the implementation is done using Tensorflow framework[4]). The grayscale input images of isolated characters are rescaled to 48x48 pixels in size. The architecture that we use (Fig. 12) 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 ReLu function and is followed by a max pooling of 2x2 blocks. The numbers of feature maps (of size 5x5) used in the three consecutive convolutional layers are 8, 16, and 32 respectively. The output of the last layers is flattened, and a fully-connected layer with 1024 neurons (also activated with ReLu) is added followed by the last output layer (softmax activation) consisting of 106 neurons corresponding to all character classes. The choices of hyper-parameters such as the number of layers and the number of feature maps are made empirically.

To avoid overfitting, dropout with probability p=0.5 is applied before the output layer. The network is trained using Adam optimizer [11] with a batch size of 100, and the training is run for 50,000 iterations after which it converges. The network produces an error rate of 6.04% on the test set.

[4] https://www.tensorflow.org

```
<CharAnno>
    <Char id="0" label="ច" lineid="0">
      <Poly x="406" y="100"/>
      <Poly x="406" y="87"/>
      ...
    </Char>
    ...
</CharAnno>
<WordAnno>
    <Word id="0" label="កំណាំង" label2="កង្កាំង">
      <CharInWord id="329"/>
      <CharInWord id="330"/>
      ...
    </Word>
    ...
</WordAnno>
```

Figure 9: Sample of an xml file storing annotation information of a manuscript page

Figure 10: Samples of annotated character patch images

Figure 11: Samples of annotated word patch images

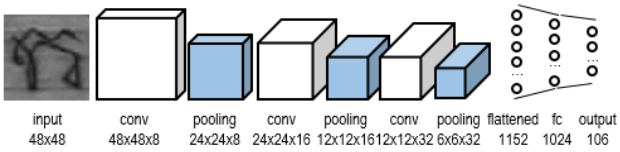

Figure 12: Architecture of the multilayer CNN

Table 4: Summary of Statistics of SleukRith Set

Nº	Data	Quantity
1	Annotated characters	301,626
2	Character classes	207
3	Annotated words	73,359
4	Unique words	6,284
5	Text lines	3,245

6 CONCLUSIONS

Datasets for ancient written character recognition algorithms are of fundamental interest for the training of statistics based recognition methods as well as for benchmarking existing recognition systems. In this paper, we introduce SleukRith Set, the first dataset constructed on digital images of Khmer palm leaf manuscripts from our own digitization campaign and also from existing digital content from various establishments. We select 657 manuscript pages, and a tool has been developed to perform the annotation task on those pages to create three types of data: isolated character dataset, annotated word dataset, and line segmentation ground truth. Furthermore, a base line experiment on isolated character recognition using a multilayer convolutional neural network is conducted. The network performs with an error rate of 6.04% demonstrating that there is still room for improvement. In future work, the next version of the dataset is likely to include an increased number of pages so that it can be used as training data for a more complex system such as deep learning. The dataset and also the annotation tool are made publicly available at github.com/donavaly/SleukRith-Set.

ACKNOWLEDGMENTS

The authors would like to thank the National Library of Cambodia, the EFEO team, and the Buddhist Institute for providing their digital images of palm leaf manuscripts. In addition, we would also like to acknowledge the help with the annotation process of our dataset by volunteer students from the Institute of Technology of Cambodia (ITC) and the National Institute of Posts, Telecommunications, and ICT (NIPTICT). This research study is supported by ARES-CCD (program AI 2014-2019) under the funding of Belgian university cooperation and the STIC Asia program implemented by the French Ministry of Foreign Affairs and International Development (MAEDI).

REFERENCES

[1] Valy, D., Verleysen, M., & Sok, K. (2016, October). Line Segmentation Approach for Ancient Palm Leaf Manuscripts Using Competitive Learning Algorithm. In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016 (pp. 108-113). IEEE.

[2] Kesiman, M. W. A., Burie, J. C., & Ogier, J. M. (2016, October). A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript. In 15th International Conference on Frontiers in Handwriting Recognition 2016, At Shenzhen, China (pp. 325-330).

[3] Chamchong, R., & Fung, C. C. (2011, September). Character segmentation from ancient palm leaf manuscripts in Thailand. In Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (pp. 140-145).

[4] Chamchong, R., & Fung, C. C. (2012, September). Text line extraction using adaptive partial projection for palm leaf manuscripts from Thailand. In International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012 (pp. 588-593). IEEE.

[5] Kesiman, M. W. A., Prum, S., Burie, J. C., & Ogier, J. M. (2016, December). Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images. In 23rd International Conference on Pattern Recognition (pp. 4006-4011).

[6] Kesiman, M. W. A., Burie, J. C., Ogier, J. M., Wibawantara, G. N. M. A., & Sunarya, I. M. G. (2016, October). AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset. In 15th International Conference on Frontiers in Handwriting Recognition 2016 (pp. 168-172).

[7] Khmer Manuscript, École Française d'Extrême-Orient. (March 2017). Retrieved from http://khmermanuscripts.efeo.fr/en/efeo-femc/project.html.

[8] Buddhist Institute. (March 2017). Retrieved from https://www.budinst.gov.kh.

[9] Natiional Library of Cambodia. (March 2017). Retrieved from https://www.facebook.com/pg/NLC.gov.kh/about.

[10] Telea, A. (2004). An image inpainting technique based on the fast marching method. Journal of graphics tools, 9(1), 23-34.

[11] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.