Image Patch Extraction in Text Area of Balinese Palm Leaf Manuscripts with Gabor Filters

Made Windu Antara Kesiman[1], Gede Aditra Pradnyana[2]
Department of Informatics Education
Universitas Pendidikan Ganesha
Singaraja Bali, Indonesia
[1]antara.kesiman@undiksha.ac.id, [2]gede.aditra@undiksha.ac.id

***Abstract*— In an effort to build a word spotting and indexation system for the Balinese palm leaf manuscripts, text area detection and image patch extraction in the text area must be done effectively. However, there is no space between words in writing with Balinese scripts. This paper aimed at describing a complete scheme to detect the text area in a Balinese palm-leaf manuscript image and to extract all possible image patches in the manuscript. Gabor filters provide initial information about the document's of text textures. A sliding window algorithm is proposed and is optimized to be able to optimally extract the image patches only in the text line area of the Gabor filtered images. The results show that the combination of Gabor filter with the optimized sliding window algorithm is effectively able to detect and to extract image patches from the text area of the Balinese palm leaf manuscripts.**

Keywords—image patch extraction; text area detection; Gabor filter; palm leaf manuscript; Balinese script.

I. INTRODUCTION

The collections of palm leaf manuscript, or Lontar, in Bali, Indonesia, record much important information and knowledge about ancient Balinese ways of life, including religion, culture, art and many other important aspects of the Balinese local wisdom. It is estimated that there are thousands collection of Lontar that can be found in Bali, including the collection in museums, cultural agencies, and also the private collections of many Balinese families at their own houses.

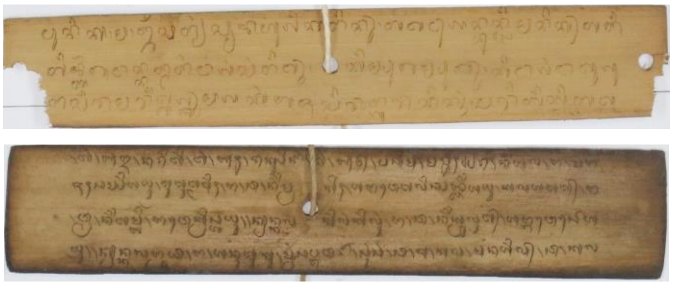


Fig. 1. Two samples of degraded manuscript pages from Bali (*Lontar*)

Unfortunately, a very limited access to the collections of Lontar make it is almost impossible to transfer and to learn the contents of manuscript. This limited access is due to two reasons. Firstly, the degraded physical condition of the manuscripts which was written on a natural dried palm leaf (Fig. 1). As what is shown in Fig. 1, there is a degradation of the manuscript because palm leaf is brittle as it is getting old. This condition makes the quality of the manuscript decrease because it is possible that the manuscript will be unreadable. Secondly, the difficulty to read the Balinese script on Lontar with a barrier of mixed old Balinese language, old Javanese language of Kawi, and the Sanskrit from India. In Bali, there are now only a few people who are able to read the Balinese script and to understand very well the contents of Lontar.

To deal with those problems, a preliminary project to digitize the collections of Lontar and to build an automatic transliteration machine to convert the Balinese script of Lontar into Roman script has been done [1], [2]. These previous works were proposed to support and to complete the final goal of building an automatic indexing system for Lontar collections. To build the indexing system and search engine for Lontar, there is also an urgent need to develop a word spotting system. The word spotting system will be helpful in finding a word patch of interest in a whole manuscript page without doing any transliteration step (Fig. 2). As what can be seen in Fig. 2, a word patch can be detected the word spotting system. Therefore, a manual transliteration will not be really necessary anymore.

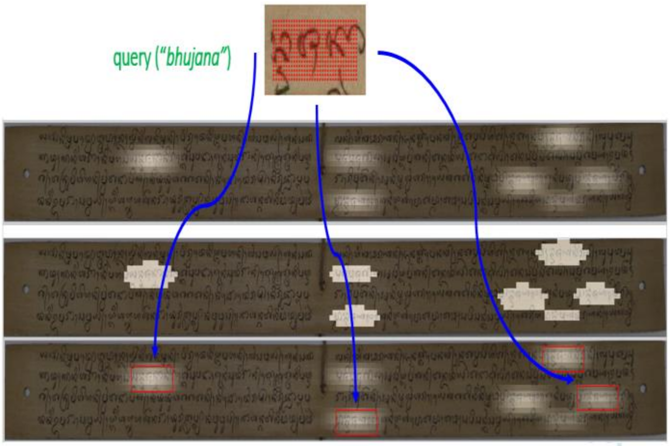


Fig. 2. The word spotting scheme for *Lontar*

This paper presents an approach to detect the text area in a Lontar page and to extract all possible word patches which can make on that page. These extracted word patches will be in the word spotting system and will be developed in future works.

Section II of this paper gives a brief description of the manuscript collections from Bali, including the Balinese script and language. The challenges in word patch extraction will also be presented in Section II. Our proposed method for image patch extraction in the text area of Lontar will be described in Section III. The experimental results and evaluation will be in Section IV. Finally, we will conclude our works with discussions for the future works in Section V.

II. THE MANUSCRIPTS FROM BALI

A. *Script and Language*

Balinese Lontar was written in the Balinese script with the Balinese language, which was mixed with Kawi language and Sanskrit. Balinese script is an alpha syllabic script where a Balinese glyph represents a syllable. The Balinese script contains more than a hundred glyphs, including basic consonant and vowel glyphs, conjunct form or second form of the consonant and vowel glyphs, and all punctuation marks and symbols [3], [4]. Two different glyphs can be written jointly in the vertical and horizontal direction with their second form glyphs as the ascender (above the medial text line) or descender (below the medial text line) glyphs (Fig.3).

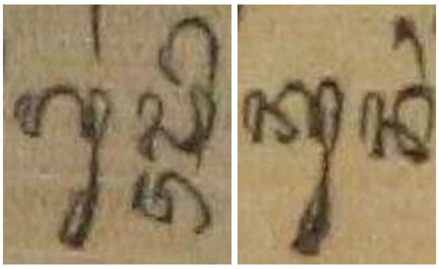


Fig. 3. Two examples of the Balinese word. Left: The word "*Gusti*" is composed of two syllables with five glyphs (two basic glyphs, one ascender glyph, and two descender glyphs). Right: word "*Kunang*" is composed of two syllables with four glyphs (two basic glyphs, one ascender glyph, and one descender glyph)

B. *Challenges in word patch extraction*

To be able to extract the possible word patches in Lontar pages, the first challenge is to detect the text area of the degraded image of Lontar with many different text layouts. Most of Lontar contain only full text (Fig. 4). It can be seen that the Lontar in fig. 4 is filled with text without any large space in it. However, in some other cases, they contain graphics with the large blank areas (Fig. 5). Fig. 5 shows that there are large blank areas which make the written text is fewer. Another kind of text is a text which is written in a table-like format (Fig. 6). The text area detection method should be able to cover all text areas and optimally avoiding the blank area to reduce the possible number of word patches that will be extracted for word spotting system. With less number of extracted word patches, the word spotting system will perform faster.

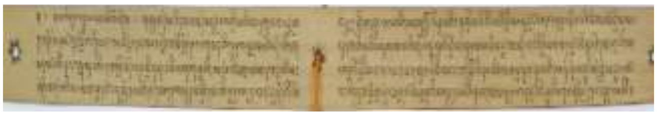


Fig. 4. Lontar page with full text

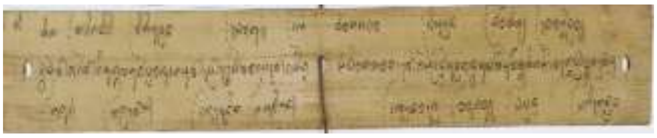


Fig. 5. Lontar page with the blank area

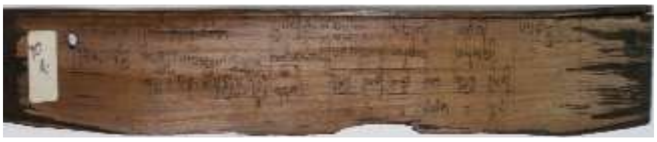


Fig. 6. Lontar page with a table-like format

The second challenge is due to the fact that there are no spaces between words in writing with the Balinese script. The word segmentation-based method cannot be trivially applied in this case [5]. Moreover, the text lines are not always written in straight position from left to right. It makes the text lines segmentation methods are difficult [5]–[10].

III. PROPOSED METHOD

Our proposed scheme of text area detection and patch image extraction for palm leaf manuscripts is a part of a global scheme for the offline patch image extraction process for the future word spotting system. Fig. 7 shows the offline patch images extraction process. Our scheme is composed of two steps: the text area detection with Gabor Filters and the patch images extraction with the Adaptive Sliding Patch Algorithm.

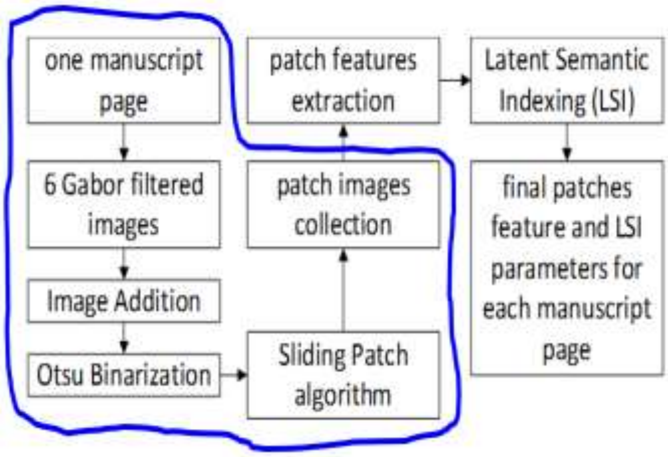

one manuscript page
6 Gabor filtered images
Image Addition
Otsu Binarization
Sliding Patch algorithm
patch images collection
patch features extraction
Latent Semantic Indexing (LSI)
final patches feature and LSI parameters for each manuscript page

Fig. 7. Offline patch images extraction process. Our proposed scheme (all steps inside the blue line) is applied before the patch feature extraction process

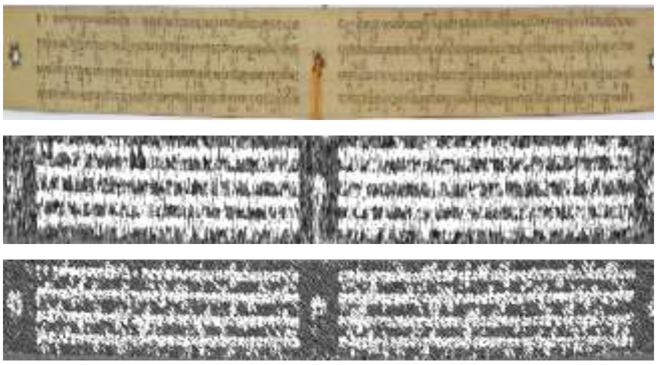
As what can be seen in fig. 7, in the process of text area detection and patch image extraction for palm leaf manuscripts, one manuscript page will be detected by gabor filters. In this process, special texture information will be detected. After that, there is a process of image addition as a combination of the gabor filtered images. Then, there is a method called Otsu Binarization. In this method, the image of text and non-text area will be clearer. Then, there is a process of sliding patch alogarithm. By the end of this process, there will be patch images collection.

A. *Text Area Detection with Gabor Filters*

Naturally, the writing in Lontar by scratching the dried palm leaf using a small-like knife shows the spatial texture information. The different writers with different writing styles represent different frequencies and orientations of the textures. To be able to detect the existence of different spatial textural areas in Lontar, a texture filter with many orientations and frequencies can be applied. One of the most commonly used textural filters is a Gabor filter. Gabor filter is a modulation between sinusoid and Gaussian filter [11]. By varying the different values of four Gabor filter parameters such as orientation, wavelength, aspect ratio, and bandwidth, we can produce a bank of Gabor filters. Each Gabor-typed filter can be used to detect different spatial texture information separately.

Gabor filters can be used to provide initial information about the existence of textures in the document. These characteristics of Gabor filters are very useful in analyzing the textural areas of the Lontar page. By investigating the different values of parameters, we can optimally detect the preliminary information about the text and no text area on Lontar.

In our works, we first extract the texture information of the gray level Lontar page image by using six different orientations of Gabor filter: 0^{o}, 45^{o}, 135^{o}, 180^{o}, 225^{o}, and 315^{o}. These orientations represent the possible scratch direction of Balinese writing for different Balinese glyphs. The horizontal orientations are not used because most of the Balinese glyph was scratched longer on vertical and diagonal direction, and also because the scratches in the horizontal direction are generally shorter and they look similar to the horizontal palm leaf textures. We set the same value for other Gabor filter parameters such as 8 for the wavelength, 0.5 for the aspect ratio, and 1 for the bandwidth (Fig. 8). We then combine the six Gabor-oriented-filtered images by using the image addition operation. The combined grayscale image will be then binarized by using the global Otsu's binarization method [12]. The resulted binary image will serve as an image mask of text and non-text area of the Lontar page (Fig. 9). The adaptive sliding patch algorithm that will be described in the next subsection will use this binary image mask as the primary input.



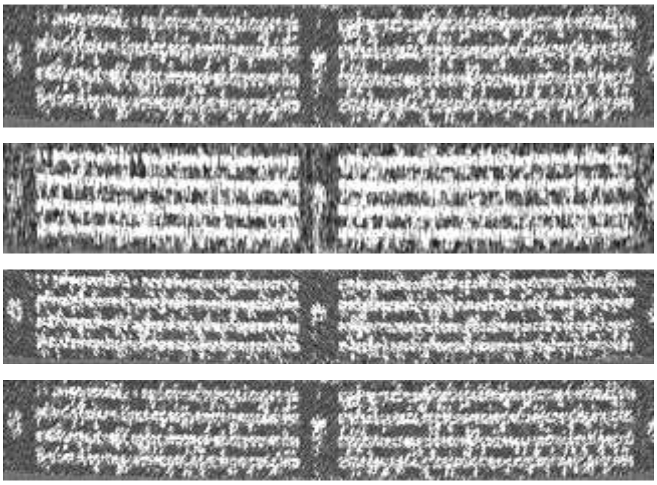


Fig. 8. The different texture orientations from six Gabor-filtered images of a manuscript page

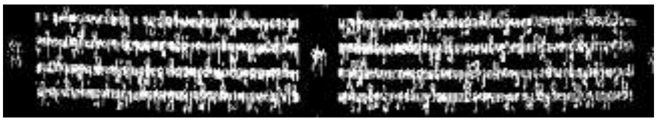


Fig. 9. The binary mask image from Fig. 8 for the adaptive sliding patch algorithm

B. *Adaptive Sliding Patch Algorithm*

By using the binary image mask, we already have the preliminary information about the text and non-text areas on Lontar pages. The basic idea of our adaptive sliding patch algorithm is to slide a small patch over the Lontar page, from left to right, from up to down, while calculating the ratio of text and non-text pixels in that patch based on the information from the binary image mask. In our approach, the sliding patch will have a more flexible possibility to slide in the vertical direction to follow the curve of the medial text lines of the Balinese writing.

In this algorithm (Algorithm 1), the size of the sliding patch is 125 (height) x 300 (width) pixels. This size is estimated based on the average size of Balinese word patches that can be found in Lontar collections. Started from the top left corner of the manuscript page, the patch will be continuously moved, first in the horizontal direction of the manuscript page, from left to right, with a step of 100 pixels. After one horizontal movement, starting from the leftmost part of the manuscript page, the patch will be moved in the vertical direction with a step of 50 pixels. During this sliding movement, the algorithm will be spotted many possible word patch images on the manuscript page.

For each word spotted patch position, the algorithm should calculate the ratio of text pixels (R). R is defined as the ratio of white pixels (text) compared to the total number of pixels in the patch image. The ratio of white pixels (text) and black pixels (non-text), but only in the upper (U), middle (M) and lower (L) part of the patch image, are also calculated to verify whether the patch image position is in the central medial text lines. These ratio values will be used as the conditional rule to decide whether this patch image is a possible word patch image, because it contains the significant number of text area (when R>0.1) and it has a good centered-spotting position on the manuscript page (when M>U and M>L).

Algorithm 1. Adaptive Sliding Patch Algorithm

```
Input: IMG_MASK : binary image as mask
       IMG      : manuscript gray scale image
Output: all possible word patch images

Algorithm:
h=125 (sliding patch height)
w=300 (sliding patch width)
h_from=1; h_to=h_from+h-1; (sliding patch height position)

while (h_from and h_to are still inside height image)
    w_from=1; w_to=w_from+w-1; (sliding patch width position)
    allow_slide=0;

    while (w_from and w_to are still inside width image AND
           h_from and h_to are still inside height image)

        Get PATCH from IMG
        Get PATCH_MASK from IMG_MASK
        Calculate R of PATCH_MASK
        Calculate U of PATCH_MASK
        Calculate M of PATCH_MASK
        Calculate L of PATCH_MASK

        if (R>0.1 and M>U and M>L) (condition is met)
            Extract this PATCH
            if (allow_slide~=0)
                h_from=h_from_save;
                h_to=h_to_save;
                allow_slide=0;
            end (of if)
        else (condition is not met)
            if (allow_slide==0) (remember the patch height position)
                h_from_save=h_from;
                h_to_save=h_to;
            end (of if)
            if (allow_slide<10) (maximum slide 10 times)
                h_from=h_from+2; h_to=h_to+2;
                w_from=w_from-w_step; w_to=w_to-w_step;
                allow_slide=allow_slide+1;
            else (no more slide, back to the saved patch height position)
                h_from=h_from_save;
                h_to=h_to_save;
                allow_slide=0;
            end (of if-else)
        end (of if-else)

        w_from=w_from+w_step; w_to=w_to+w_step;
    end (of while)
    h_from=h_from+h_step; h_to=h_to+h_step;
end (of while)
```

If the conditional rules are met, the algorithm will consider the patch image position as the good word image patch. In the future process, this patch will be sent to the patch feature extraction module. Otherwise, when the conditional rules are not met, the algorithm should find other possible word patch images by sliding the patch position to 2 pixels in a lower vertical position repetitively until the conditional rule is finally met. During this repetitive patch searching procedure in a vertical position, the horizontal position of the patch should be kept. The algorithm will be stopped when the sliding patch image already visited all parts of the Lontar page.

IV. Experimental Results and Evaluation

To see the effectiveness of the method, we performed our proposed scheme in different pages of Lontar with the different degraded conditions and different text layouts.

A. Results

By applying the text area detection with Gabor filters and the adaptive sliding patch algorithm, we can extract the possible word patch image on the text area of the Lontar page (Fig. 10 & 11).

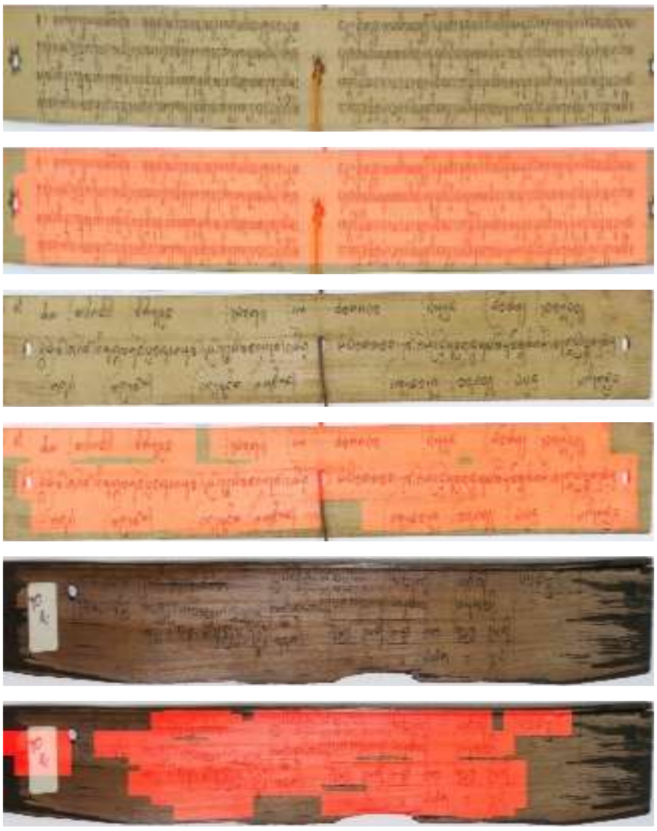


Fig. 10. Examples of text patch area extraction (in red) from different manuscript pages with the different degraded condition and text layouts

In Fig. 10, it shows that the first and the second manuscript contain only full text. The third and the forth manuscript contain graphics with large blank ares. Then, the fifth and the sixth manuscript are written in a table-like format. Those manuscripts are detected with Gabor filters and the adaptive sliding patch algorithm, so possible word patch image can be extracted. Fig. 11 shows the possible word patch image on the area of the text of the Lontar page.

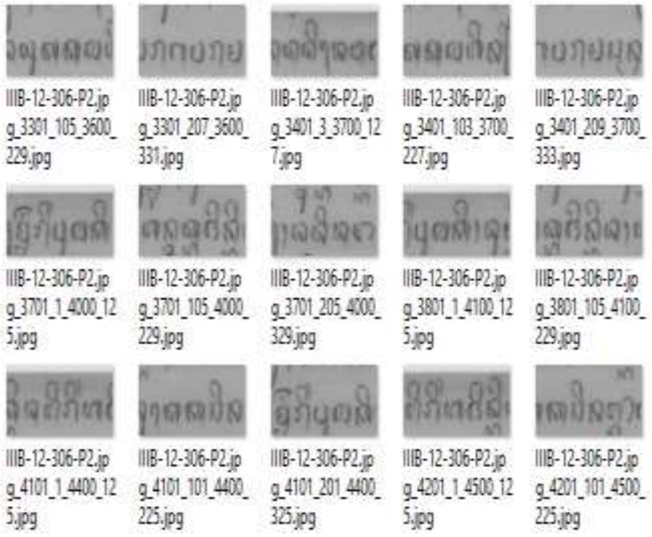

IIIB-12-306-P2.jpg_3301_105_3600_229.jpg
IIIB-12-306-P2.jpg_3301_207_3600_331.jpg
IIIB-12-306-P2.jpg_3401_3_3700_127.jpg
IIIB-12-306-P2.jpg_3401_103_3700_227.jpg
IIIB-12-306-P2.jpg_3401_209_3700_333.jpg
IIIB-12-306-P2.jpg_3701_1_4000_125.jpg
IIIB-12-306-P2.jpg_3701_105_4000_229.jpg
IIIB-12-306-P2.jpg_3701_205_4000_329.jpg
IIIB-12-306-P2.jpg_3801_1_4100_125.jpg
IIIB-12-306-P2.jpg_3801_105_4100_229.jpg
IIIB-12-306-P2.jpg_4101_1_4400_125.jpg
IIIB-12-306-P2.jpg_4101_101_4400_225.jpg
IIIB-12-306-P2.jpg_4101_201_4400_325.jpg
IIIB-12-306-P2.jpg_4201_1_4500_125.jpg
IIIB-12-306-P2.jpg_4201_101_4500_225.jpg

Fig. 11. Some examples of the good centered-spotting position from the extracted patch images of a Lontar page

In Fig. 11, it can be seen that there are some extracted patch images as the results of Gabor filters and the adaptive sliding patch algorithm, we can see that the extracted patch images look clearer despite the degradation of the palm leaf manuscrips.

B. *Evaluation*

The combination of Gabor filter with the optimized sliding window algorithm is effectively able to detect and to extract image patches from the text area of the Balinese palm leaf manuscripts. The results show visually that our proposed scheme extracts the word patch images in the optimal text area position in central-medial text lines. This procedure is well adapted to deal with curved unstraight text line in manuscript and different text and non-text layout on Lontar pages.

V. Conclusions and Future Works

We presented in this paper an approach to detect the text area in a Lontar page and to extract all possible word patches which can be found on that page. Gabor filters are used to provide initial information about the existence of text textures in the document. An algorithm with the sliding window concept is proposed and is optimized to be able to optimally extract the image patches only in the text line area of the Gabor filtered images. The results show that the combination of Gabor filter with the optimized sliding window algorithm is effectively able to detect and to extract image patches from the text area of the Balinese palm leaf manuscripts. These extracted word patches will be used in the word spotting system that will be developed in future works.

Acknowledgment

The authors thank Museum Gedong Kirtya and Museum Bali, the Balinese families and students of Universitas Pendidikan Ganesha who provide the dataset and help in collecting and ground-truthing the dataset for this work. This research is supported by RISTEK DIKTI PDUPT 2019 Research Funding Program and from the French government, the STIC Asia Program (2015-2018).

References

[1] M. Kesiman *et al.*, “Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia,” *Journal of Imaging*, vol. 4, no. 2, p. 43, Feb. 2018.

[2] M. W. A. Kesiman, J.-C. Burie, and J.-M. Ogier, “A Complete Scheme Of Spatially Categorized Glyph Recognition For The Transliteration Of Balinese Palm Leaf Manuscripts,” in *14th IAPR International Conference on Document Analysis and Recognition*, Kyoto, Japan, 2017.

[3] M. W. A. Kesiman, S. Prum, J.-C. Burie, and J.-M. Ogier, “Study on Feature Extraction Methods for Character Recognition of Balinese Script on Palm Leaf Manuscript Images,” in *23rd International Conference on Pattern Recognition*, Cancun, Mexico, 2016.

[4] M. W. A. Kesiman, J.-C. Burie, J.-M. Ogier, G. N. M. A. Wibawantara, and I. M. G. Sunarya, “AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset,” in *15th International Conference on Frontiers in Handwriting Recognition 2016*, Shenzhen, China, 2016, pp. 168–172.

[5] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, “Text line and word segmentation of handwritten documents,” *Pattern Recognition*, vol. 42, no. 12, pp. 3169–3183, Dec. 2009.

[6] N. Arvanitopoulos and S. Susstrunk, “Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts,” in *2014 14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 726–731.

[7] A. Asi, R. Saabni, and J. El-Sana, “Text line segmentation for gray scale historical document images,” in *HIP ’11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, 2011, p. 120.

[8] J.-C. Burie *et al.*, “ICFHR 2016 Competition on the Analysis of Handwritten Text in Images of Balinese Palm Leaf Manuscripts,” in *15th International Conference on Frontiers in Handwriting Recognition 2016*, Shenzhen, China, 2016, pp. 596–601.

[9] M. W. A. Kesiman, J.-C. Burie, and J.-M. Ogier, “A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript,” in *15th International Conference on Frontiers in Handwriting Recognition 2016*, Shenzhen, China, pp. 325–330.

[10] M. W. A. Kesiman *et al.*, “Southeast Asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges,” *Journal of Electronic Imaging*, vol. 26, no. 1, p. 011011, Nov. 2016.

[11] V. Shiv Naga Prasad and J. Domke, “Gabor filter visualization,” University of Maryland, Technical Report, 2005.

[12] “Global image threshold using Otsu’s method - MATLAB graythresh - MathWorks France.” [Online]. Available: https://fr.mathworks.com/help/images/ref/graythresh.html?requestedDomain=true. [Accessed: 20-Feb-2018].