Presented at 16th ICFHR 2018, Niagara Falls, USA, August 5-8, 2018

Training Schemes for the Transliteration of the Balinese Script into the Latin Script on Palm Leaf Manuscript Images

Made Windu Antara Kesiman[1], Jean-Christophe Burie[1], Jean-Marc Ogier[1], and Philippe Grangé[2]

[1]Laboratoire Informatique Image Interaction (L3i), [2]Department of Indonesian Language, Université de La Rochelle, France

Research context

- ✓ Considering the importance of the contents of the Balinese palm leaf manuscripts, transliteration system has to be developed in order to be able to read easily these manuscripts.
- ✓ The challenge comes from the fact that Balinese script is a syllabic script.
- ✓ With a very limited training data availability, some adaptations in the transliteration training scheme need to be designed, to be analyzed and to be evaluated.

Challenges in Text Transliteration

- ✓ The problem of one-to-one mapping between linguistic symbols and images of symbols
- ✓ Number of combination of possible compound syllable will be huge, and collecting enough labeled samples for each class is hard and it needs an extraordinary effort.
- ✓ The problem of allographs [3], where more than one shape of glyph (image of symbol) is allowed to be used to represent a same sound of speech of syllable (linguistic symbol).

Corpus

- ➢ Sample images of the palm leaf manuscripts from Bali, Indonesia, from 23 different collections, from 5 different locations (regions): 2 museums and 3 private families [1].
- ➢ Images of manuscript: 303 pages from 22 collections with 1172 total text lines are used.
- ➢ Transliterated Latin text of the manuscript corpus: 8,662 unique real words.
- ➢ Isolated real glyph annotated images: 19,383 real glyph annotated images from 133 classes.
- ➢ Real word annotated images: 15,022 real word images from 130 pages for training and 10,475 real word images from 100 pages for testing.

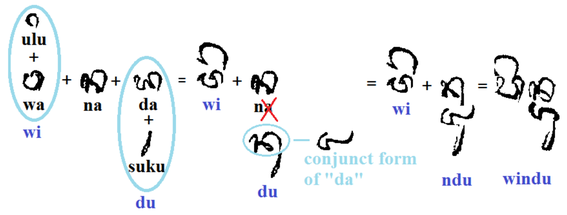
Training Scheme at Word Level

Scheme W1: real word image samples from real word annotated images.
Scheme W2: meaningful synthetic word image samples generated from real words (in the corpus).
Scheme W3: meaningful synthetic word image samples generated from real words (not in the corpus).

Training Scheme at Text Line Level

Scheme T3: meaningful synthetic text line image samples generated from real words (not in the corpus) and with spaces between words.
Scheme T4: meaningful synthetic text line image samples generated from real words (not in the corpus) and without any spaces between words.
Scheme WT (Word-Textline): meaningful synthetic text line image samples generated from real sentences (not in the corpus). The pre-trained network from word level of Scheme W1 is used.

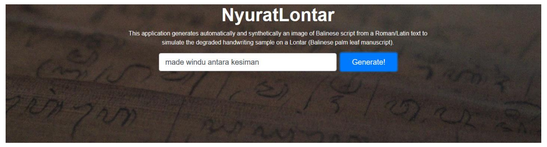
[1] M W A Kesiman, J-C Burie, J-M Ogier, G N M A Wibawantara, and I M G Sunarya. AMADI lontarset: The First Handwritten Balinese Palm Leaf Manuscripts Dataset. In 15th International Conference on Frontiers in Handwriting Recognition 2016, pages 168–172, Shenzhen, China, October 2016.
[2] https://github.com/tmbdev/ocropy
[3] David Doermann and Karl Tombre, editors. Handbook of Document Image Processing and Recognition. Springer London, London, 2014.
[4] 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, November 2017.



Allographs

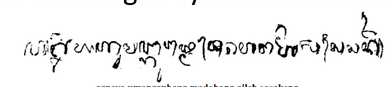
Automatic Synthetic Handwritten Balinese Script Generator

- ❑ Generates automatically and synthetically an image of Balinese script
- ❑ Input: Latin text, Output: degraded handwriting sample on a Balinese palm leaf manuscript
- ❑ Used to render the meaningful real words and text lines into the meaningful synthetic Balinese script images



Here, the Balinese script images are generated from almost 20,000 Balinese glyph samples from palm leaf manuscript collections. Some of the glyphs are still missing, we will try to complete our glyph samples from other palm leaf manuscript collections for the next version of this application.

"made windu antara kesiman" in Balinese script:



agneya umanguhang madohang aileh sasakeng

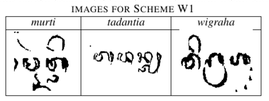
Fig. 3. Meaningful synthetic text line image samples generated from real words (not in the corpus) and with spaces between words for Scheme T3

kasaambi,Langetanget,buanajiwatem,atmawisesahadaang

Fig. 4. Meaningful synthetic text line image samples generated from real words (not in the corpus) and without any spaces between words for Scheme T4

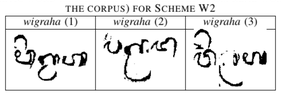
Real word image samples collected from real word annotated images for Scheme W1

murti	*tadantia*	*wigraha*



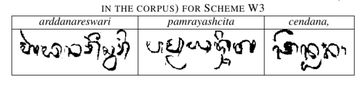
Synthetic word image samples generated from real words (in the corpus) for Scheme W2

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



Synthetic word image samples generated from real words (not in the corpus) for Scheme W3

arddanareswari	*pamrayashcita*	*cendana.*



Experimental Protocols

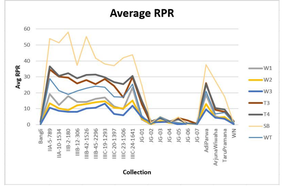
- The OCRopy framework [2] : using RNN-LSTM architecture, the sequence depth of 100 pixels, the neuron size of 200
- The segmentation based transliteration method (Scheme SB) [4] will also be tested and evaluated as comparison.

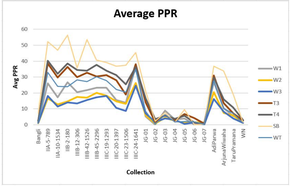
Evaluation Metrics

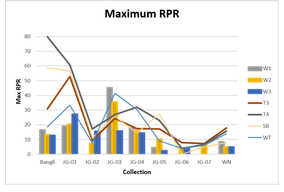
- ❖ For word transliteration: Character Error Rate (CER)
- ❖ For text line transliteration: Recall Pattern Rate (RPR) and Precision Pattern Rate (PPR)

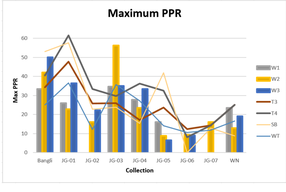
Error Rate of Word Transliteration

Scheme	W1	W2	W3	T3	T4	WT	SB
CER (%)	39.70	60.24	63.45	64.52	64.64	62.31	57.06









Word transliteration: training schemes at word level perform better than training schemes at text line level. Text line transliteration: segmentation based method outperforms all segmentation free training schemes for the less degraded collections, while the segmentation free training schemes contributes in transliterating the more degraded manuscripts.