The Challenges of Image Analysis, Recognition and Understanding on Old Sundanese Palm Leaf Manuscript

Erick Paulus

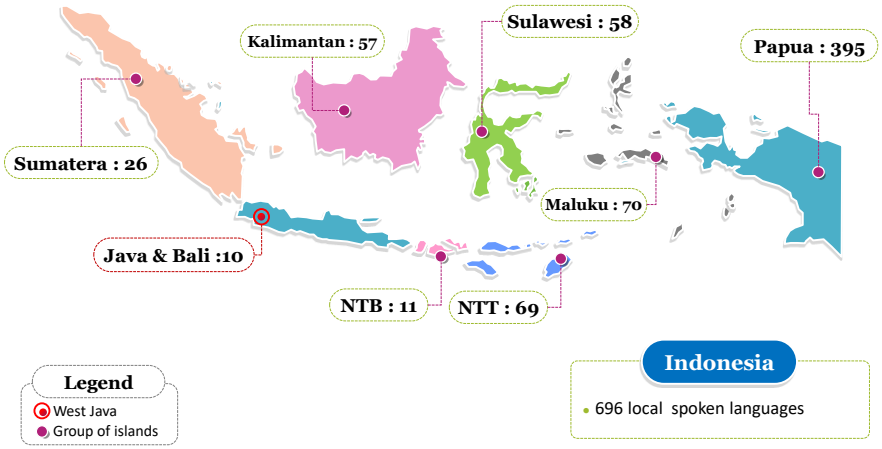
Processing Ancient Text Corpora
@Lorentz Center , 20 Feb 2020

Universiteit Leiden
The Netherlands

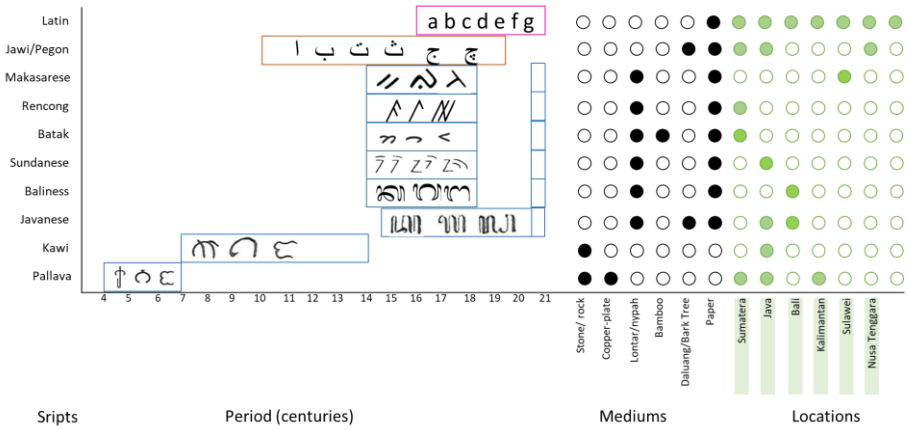
Supervised by
Prof. Fons J. Verbeek (LIACS)
Prof. J.C Burie (La Rochele University, France)

Background

- **Spoken Language vs. Writing System**



Writing System Mapping



Background

- **Language vs. Writing System**
- **Old Sundanese palm leaf manuscripts,**
  - written in XIV-XVIII centuries
  - stores the knowledge in religious value, sciences, medicine, literature, and so on.
  - **Old Sundanese script** and Buda Script

Context of Palm leaf manuscripts

Firstly

- ❑ used in India (Pallava Script)

Then

- ❑ adopted in southeast Asia (Kawi, old Javanese, old Sundanese, old Balinese, Buda, khemr script, and so on…

Context of Palm leaf manuscripts

- ❑ Characteristics

Normally, in one leaf, there are 2-4 text lines

Number of leaves for a collection is not constant :

- For the collection of Ramayana (Putra Rama dan Rawan)-Koropak 22 = 36 leaves
- For the collection of Kakawin Ramayana – koropak 21 = 13 leaves

Purpose of the research project

❑Provide new tools to access to the content
- Quickly
- Efficiently

❑Potential users
- Philologist, Linguist, Historian
- All people interested in the content of the palm leaf manuscripts

❑Main issues
- Old documents (sometimes damaged)
- Specific material (palm leaves)
- Old languages (difficulties to find expert)
- Handwritten text (variability of the writing)

Issues on Document Image Analysis

- Digitization campaign
- Annotation and label -> generating the ground truth
- Binarization and enhancement of the quality of image
- Segmentation (binarization-free vs binarization-based)
- Feature Extraction
- Classification (Trained vs handcrafted feature)
- Transliteration (knowledge representation and phonological rules)
- Writer verification and Identification

Old Sundanese Manuscript Collections

Location	Type of writing medium	Sum of collections	Sum of pages
Kabuyutan Ciburuy (private family collection) [1]	Palm leaf (27 collections),	27	1452
National Library of Indonesia (PNRI) [2]	Palm leaf (33 collections), Nypa (20 collections), Bamboo (3 collections), and Daluwang paper (2 collections)	58	?
Leiden University Library		?	?

[1]Result of acquisition process
[2]M. Holil and A. Gunawan, "Membuka Peti Naskah Sunda Kuna Koleksi Perpustakaan Nasional RI: Upaya Rekatalogisasi," in *Simposium Internasional Pernaskahan Nusantara XIII*, 2010.

Image Digitization

No	Location/Source	Collections	Pages	Digitization Tools
1	Kabuyutan Ciburuy	27	1452	Canon EOS 5D Mark II

[1]Result of acquisition process
[2]M. Holil and A. Gunawan, "Membuka Peti Naskah Sunda Kuna Koleksi Perpustakaan Nasional RI: Upaya Rekatalogisasi," in *Simposium Internasional Pernaskahan Nusantara XIII*, 2010.

Socio-Cultural Challenges

- Difficulty in collecting Data
  - Cultural and religious rules
  - Sacred collections
  - Permission
- Difficulty in finding the Sundanese philologist
  - Never used since 18th centuries
  - Not popular
  - Most Sundanese can not read
  - Retirement
  - Lack experience

How digitization is processed?

- The lontars are put out from the box
- Every lontar is cleaned from the dust
- smeared with Kemiri oil (candle nut oil) to deal with fading script
- and wiped with the tissue
- Take a photo
- Lontar is put back into the box in order

The sample images from Ciburuy Collections

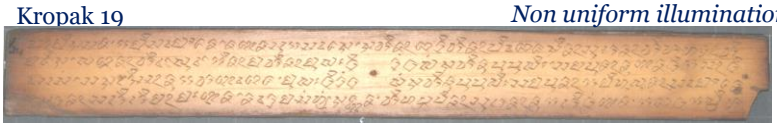
Ciburuy X

Ciburuy XI

Ciburuy XV

The sample images from Ciburuy Collections

Kropak 19 *Non uniform illumination*



Ciburuy VI *fading*

Ciburuy VII *Smudge*

The sample images from Ciburuy Collections

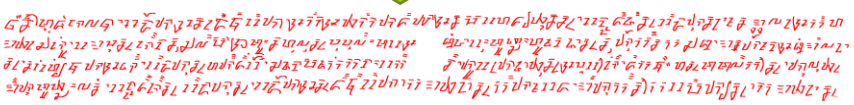
Ciburuy II

broken

Generating ground truth images

- The generating of ground truth of historical manuscript using PixLabeler.
- Done by philologist and students from philology department
- 1 lontar : 2-4 hours (depending on the sum of glyphs)

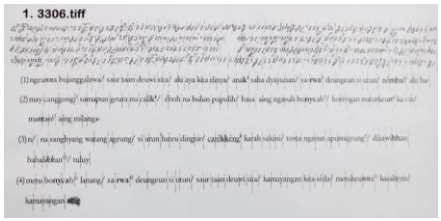
Binarization using brush technique in PixLabeler



The result of binarization process

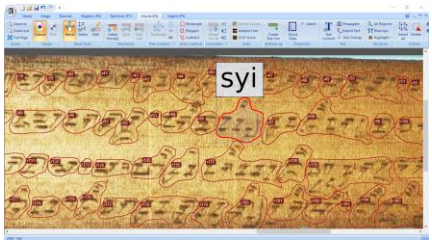
Annotation in Syllable Level

- The manual segmentation and annotation by the **philologist** transfer into digital segmentation and annotation using Aletheia



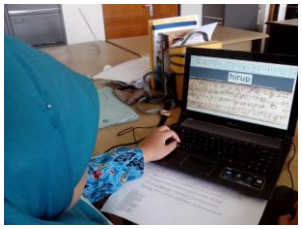
Manual segmentation and annotation by philologist

Segmentation and annotation in syllable level using Aletheia



Annotation in Word Level

- For manual process : Time consuming, costly,
- 1 lontar : 2-4 hours
- The results from segmented and annotated manually by the philology transfer into digital segmentation and annotation using Aletheia



The segmentation and annotation in word level using Aletheia

Sundanese Set

- Summary of Statistics

No	Data	Quantity
1	Collections	5
2	Pages	66
3	Annotated Characters/Glyphs	7371
4	Character Classes	60
5	Annotated Words	1843
6	Text Lines	242

Sundanese Dataset https://goo.gl/EU41gt

Experimental results for binarization task

Method	Parameter	FM(%)	NRM	PSNR
Otsu Gray		23.70566	0.326681	9.998433
Sauvola	window = 50, k = 0.2, R = 128	43.04994	0.299694	23.65228
Niblack	window = 50, k = −0.2	46.79678	0.195015	20.31759
NICK	window = 50, k= −0.2	29.5918	0.390431	24.26187
Howe	Default Value	45.90779	0.235175	21.90439

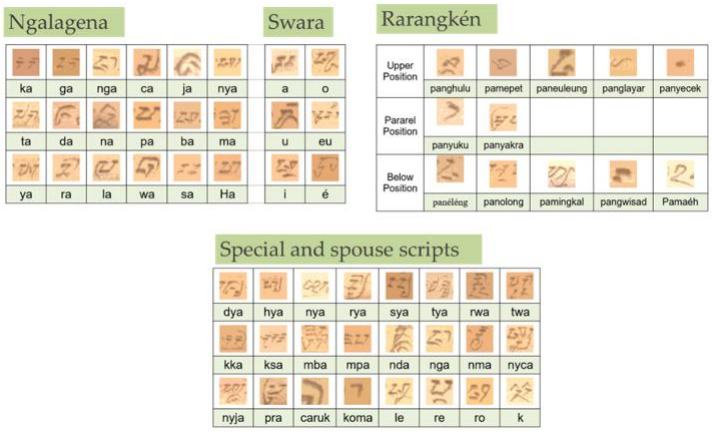
Experimental results for text line segmentation task

	N	M	o20	DR	RA	FM
Seam carving	46	43	36	78.26	83.72	80.89
	242	257	218	90.08	84.82	87.37
Adaptive Path Finding	46	50	41	89.13	82	85.41
	242	253	222	91.73	87.74	89.69

the count of ground truth elements (N),
the count of result elements (M),
the one-to-one (o20) match score is computed for a region pair based on 90% acceptance threshold,
detection rate (DR),
recognition accuracy (RA), and
Performance metric (FM).

21

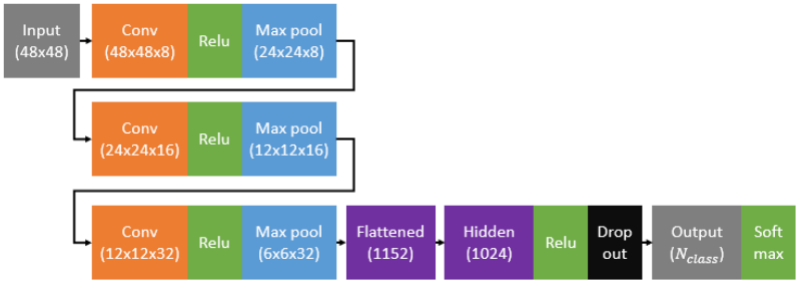
Dataset- 60 classes of old Sundanese glyph



Training : 4555 isolated glyph images
Testing : 2816 isolated glyph images

CNN Architecture

- It achieved to 79.05 % recognition rate



So, these are the plans

- To generate a new dataset by collecting 200 pages of ancient Sundanese manuscripts from different writers stored in Leiden University Library
- To improve the binarization strategies for enhancing the quality of image
- To improve classification strategies (character level and word level)
- To build keyword spotting (machine learning approaches)
- To construct the transliteration role or phonological role for old Sundanese script

Acknowledgement

- **Supervisor**
  - Prof. Fons J. Verbeek (LIACS)
  - Prof. Jean-Christophe Burie (La Rochele Univ, France)
- **Partner**
  - Rahmat Sopian, M.Hum (a philologist and Ph.D Candidate from Tokyo University of Foreign Studies – supervised by Prof. Toru Aoyama, Ph.D)
  - Dr. Undang Ahmad Darsa, M.Hum. (as senior philologist from Padjadajran University)
  - Riki Nawawi, S.Hum. (as junior philologist and Master Student from Padjadajran University)
  - Dr. Setiawan Hadi, M.Sc.CS. (A research leader of Ancient Manuscript Digitization and Indexation from Padjadajran University)
  - Prof. Dr. Budi Nurani Ruchjana (Head of Center of Etnoscience Studies from Padjadajran University)
  - Prof. Benhard Arps (Professor of Indonesian and Javanese Language and Culture from Leiden Univ)

Thank you

Universiteit Leiden
The Netherlands