technische universität dortmund

fakultät für informatik

Lampung - a New Handwritten Character Benchmark: Database, Labeling and Recognition

Akmal Junaidi, Szilárd Vajda, Gernot A. Fink

Computer Science Department, TU Dortmund, Germany

{*akmal.junaidi,szilard.vajda,gernot.fink*}*@udo.edu*

September 17, 2011

Overview of the talk:

- Introduction
  - Motivation
  - Script
- Labeling
- Features
- Experiments
- Conclusion

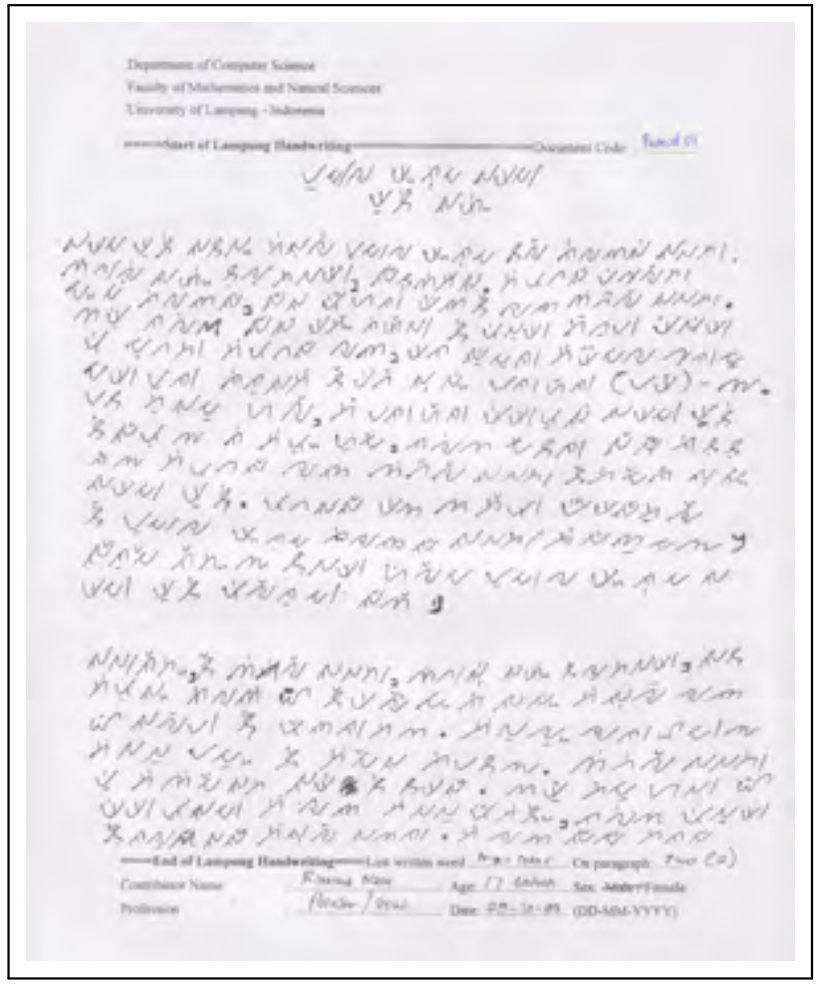
Motivation

New script:

- lack of publications
- no representative dataset

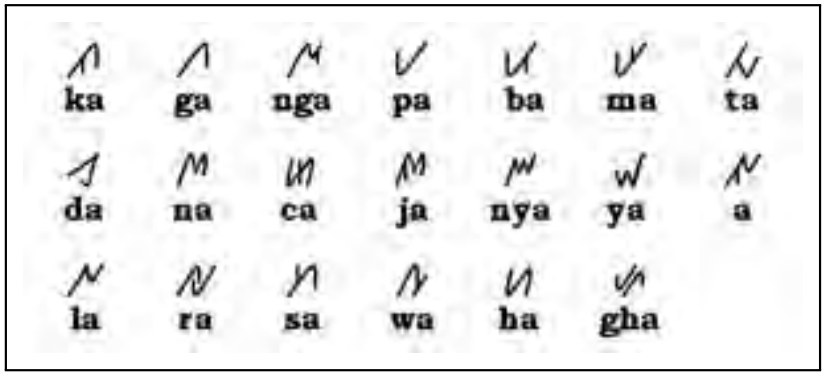
Cultural heritage

- originated from Brahmi script
- preserving important heritage
- proof of script existence



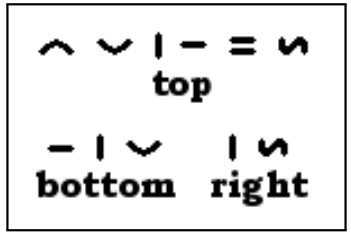
Lampung alphabet

Characteristics:

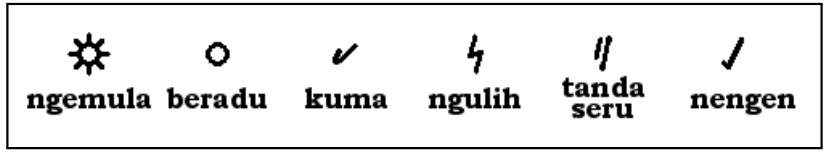


- not cursive
- curve(s)
- 20 letters
- the name: Kaganga

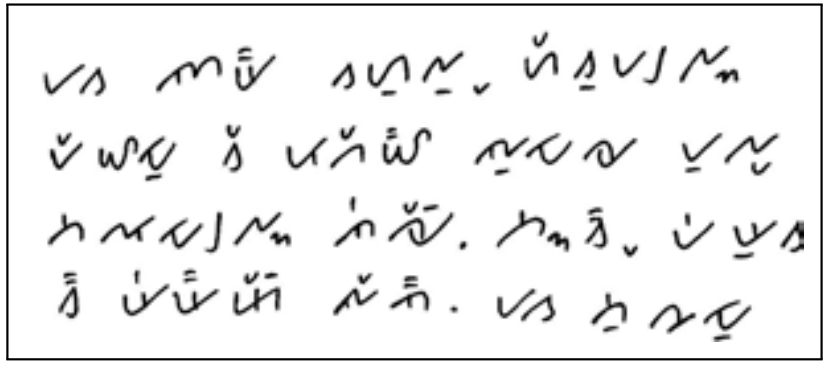
Diacritics:



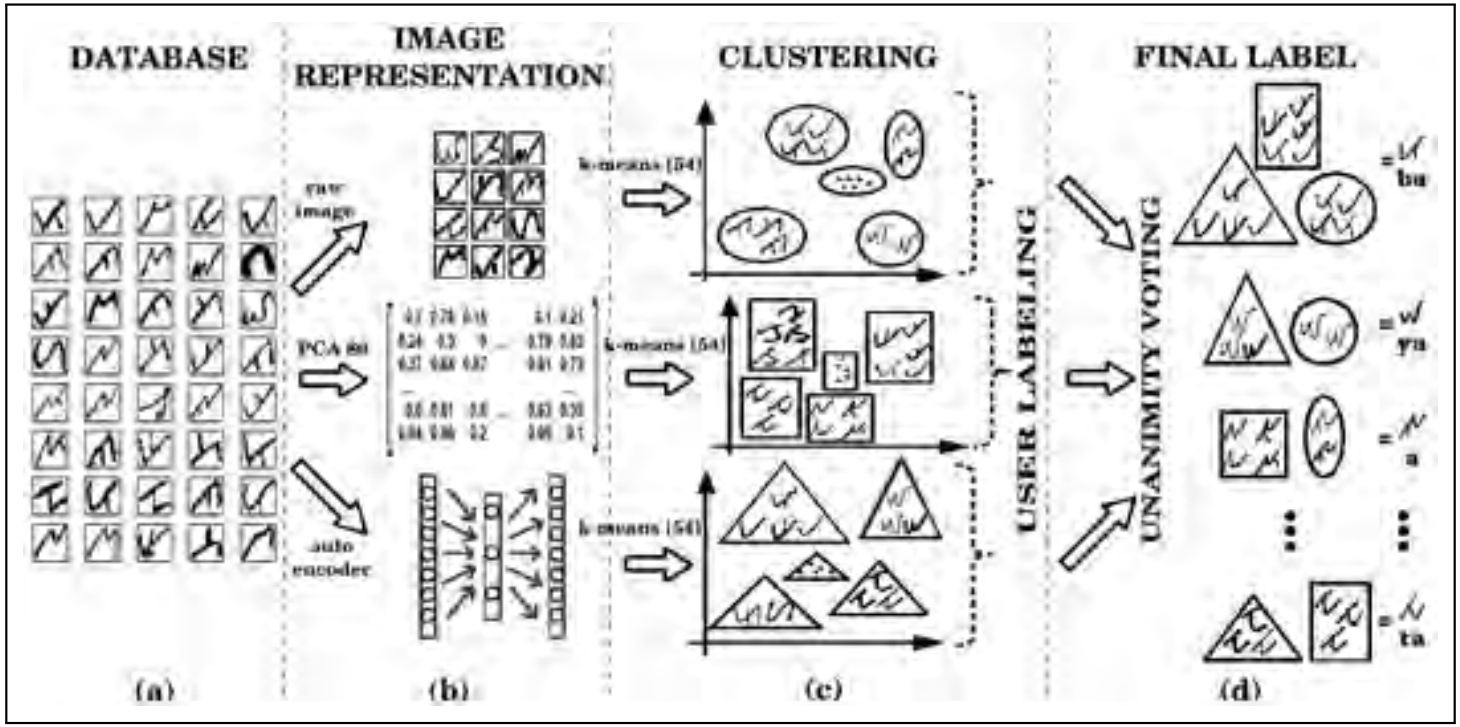
Punctuation marks



Handwriting sample



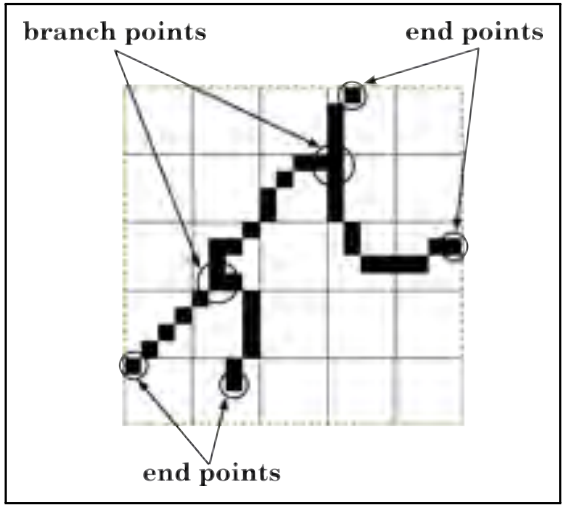
Semi-Automatic Labeling: An overview [1]



[1] Vajda et.al, Semi-Supervised Ensemble Learning Approach for Character Labeling with Minimal Human Effort, ICDAR, 2011

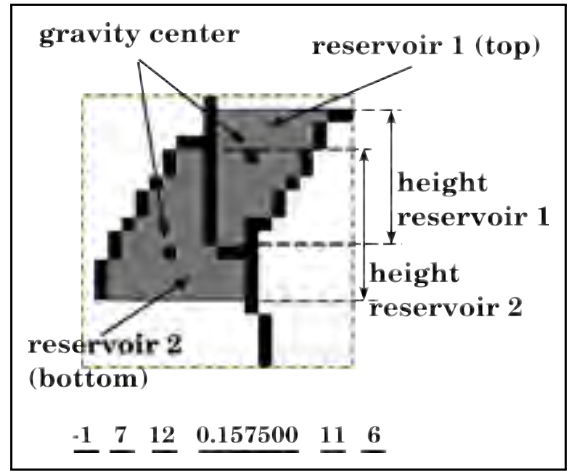
Features

Structural and statistical:



- branch points
- end points
- pixel density

Water reservoir:



- top and bottom
- gravity center
- size (volume)
- height and width

Experiments

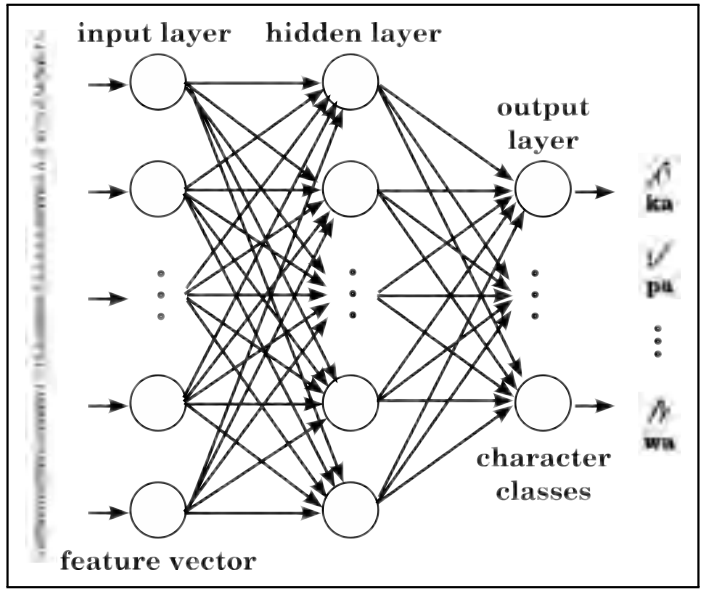
Dataset:

- fairy tales transcription
- 82 docs. written by students
- 35,193 character images
- clustered to 11 classes

Composition:

- 21,122 for training set (60%)
- 10,547 for test set (30%)
- 3,524 for validation set (10%)

Classification: Neural network



Recognition result

Features	#Training	#Test	Rec (%)
Branch points, end points, pixel density (BED)	21,122	10,547	93.2±0.48
Water reservoirs (WR)	21,122	10,547	91.3±0.54
BED and WR	21,122	10,547	94.3±0.44

Misclassification

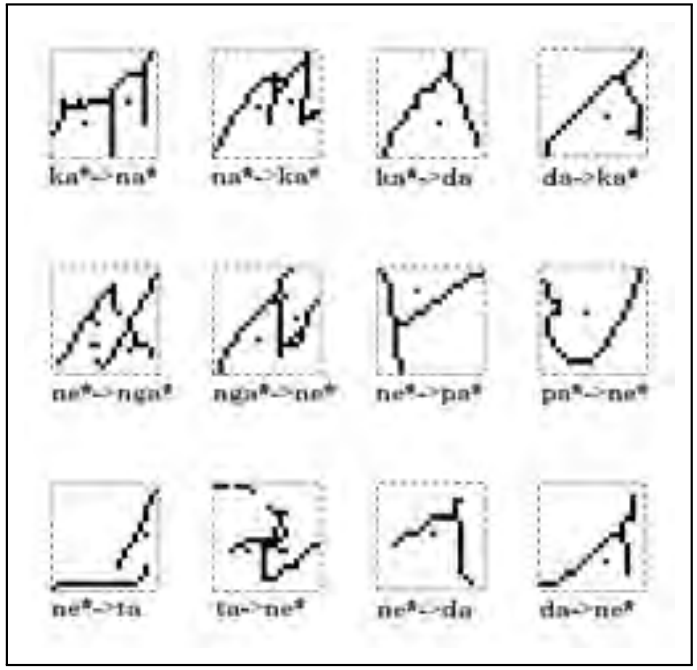
Variability in writing style

Different location of water reservoir

Unfiltered punctuation marks

Artifacts:

- touching characters
- character connected to diacritic(s)
- character connected to punctuation mark(s)



Conclusion

- The Lampung:
  - scientific research challenge for handwritten recognition
  - preserving efforts of the Lampung as a cultural heritage
- Semi-automatic labeling strategy: new approach
  - efficient labeling task for large dataset, minimize human involvement
  - only 20% samples need to be relabeled
- Water reservoir can effectively distinguish the Lampung characters:
  - 91.3% recognition only based on water reservoir features
  - 94.3% recognition combining with branch points, end points, pixel density
- Lampung character dataset:
  - publicly available soon
  - preferably on TC11 website

References I

[1] U. Bhattacharya and B. B. Chaudhuri.
Databases for Research on Recognition of Handwritten Characters of indian Scripts.
In *International Conference on Document Analysis and Recognition*, volume 2, pages 789 – 793, 2005.

[2] B. B. Chaudhuri and S. Ghosh.
Orientation Detection of Major Indian Scripts.
In *Proceedings of the International Workshop on Multilingual OCR*, MOCR '09, pages 8:1–8:7, New York, NY, USA, 2009. ACM.

[3] P. T. Daniels.
The World's Writing Systems.
Oxford University Press, 1996.

[4] D. Ghosh, T. Dube, and A. Shivaprasad.
Script Recognition: A Review.
IEEE Trans. Pattern Anal. Mach. Intell., 32:2142–2161, December 2010.

[5] G. E. Hinton and R. R. Salakhutdinov.
Reducing the Dimensionality of Data with Neural Networks.
Science, 313(5786):504–507, July 2006.

[6] M. S. Khorsheed.
Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model.
Pattern Recogn. Lett., 24:2235–2242, October 2003.

[7] L. I. Kuncheva.
Combining Pattern Classifiers: Methods and Algorithms.
Wiley-Interscience, 2004.

References II

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.
Gradient-Based Learning Applied to Document Recognition.
In *Intelligent Signal Processing*, pages 306–351. IEEE Press, 2001.

[9] C.-L. Liu and C. Y. Suen.
A New Benchmark on the Recognition of Handwritten Bangla and Farsi Numeral Characters.
Pattern Recognition, 42:3287–3295, December 2009.

[10] L. M. Lorigo and V. Govindaraju.
Offline Arabic Handwriting Recognition: A Survey.
IEEE Trans. Pattern Anal. Mach. Intell., 28:712–724, May 2006.

[11] T. Mondal, U. Bhattacharya, S. K. Parui, K. Das, and V. Roy.
Database Generation and Recognition of Online Handwritten Bangla Characters.
In *Proceedings of the International Workshop on Multilingual OCR*, MOCR '09, pages 9:1–9:6, New York, NY, USA, 2009. ACM.

[12] S. Mozaffari, H. E. Abed, V. Märgner, K. Faez, and A. Amirshahi.
IfN/Farsi-Database: a Database of Farsi Handwritten City Names.
In *International Conference on Frontiers in Handwriting Recognition*, 2008.

[13] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban, and S. M. Golzan.
A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research.
In *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France), 2006.

[14] W. Niblack.
An Introduction to Digital Image Processing.
Strandberg Publishing Company, Birkeroed, Denmark, 1985.

References III

[15] U. Pal, A. Belaïd, and C. Choisy.
Touching Numeral Segmentation using Water Reservoir Concept.
Pattern Recognition Letters, 24(1-3):261–272, 2003.

[16] U. Pal and S. Datta.
Segmentation of Bangla Unconstrained Handwritten Text.
In *International Conference on Document Analysis and Recognition*, pages 1128–1132, 2003.

[17] U. Pal, S. Kundu, Y. Ali, H. Islam, and N. Tripathy.
Recognition of Unconstrained Malayalam Handwritten Numeral.
In *ICVGIP*, pages 423–428, 2004.

[18] U. Pal, R. K. Roy, K. Roy, and F. Kimura.
Indian Multi-Script Full Pin-code String Recognition for Postal Automation.
In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 456–460, Washington, DC, USA, 2009. IEEE Computer Society.

[19] T. Pudjiastuti.
The Lampung Ancient Script and Manuscript in Perspective of the Recent Contemporary Lampung Society (Indonesian).
Cultural and Education Department, Republik of Indonesia, Jakarta, 1997.

[20] P. P. Roy, U. Pal, and J. Lladós.
Morphology Based Handwritten Line Segmentation Using Foreground and Background Information.
In *International Conference on Frontiers in Handwriting Recognition*, 2008.

References IV

[21] N. Stamatopoulos, G. Louloudis, and B. Gatos.
Efficient Transcript Mapping to Ease the Creation of Document Image Segmentation Ground Truth with Text-Image Alignment.
In *International Conference on Frontiers in Handwriting Recognition*, pages 226–231, Washington, DC, USA, 2010. IEEE Computer Society.

[22] S. Vajda and G. Fink.
Exploring Pattern Selection Strategies for Fast Neural Network Training.
In *International Conference on Pattern Recognition*, pages 2913 –2916, 2010.

[23] S. Vajda, A. Junaidi, and G. A. Fink.
A Semi-Supervised Ensemble Learning Approach for Character Labeling with Minimal Human Effort.
In *International Conference on Document Analysis and Recognition*, 2011.
(in press).

[24] S. Vajda, T. Plötz, and G. A. Fink.
Layout Analysis for Camera-Based Whiteboard Notes.
Journal of Universal Computer Science, 15(18):3307–3324, 2009.

[25] S. Vajda, K. Roy, U. Pal, B. B. Chaudhuri, and A. Belaid.
Automation of Indian Postal Documents Written in Bangla and English,.
International Journal of Pattern Recognition and Artificial Intelligence, 23(8):1599–1632, December 2009.