Recognizing Vietnamese online handwritten separated characters

Duy K.N., The D.B.
College of Technology, Vietnam National University, Hanoi

Abstract: Vietnamese alphabet is based on the Latin alphabet with the addition of nine accent marks or diacritics four of them to create additional sounds, and the other five to indicate the tone of each word. Because Vietnamese is a tonal language that uses tone to distinguish words, recognizing diacritics is an important part in recognizing Vietnamese word. However, in written form, diacritics are much smaller than the characters, which make very them hard to recognize. Previous works on Vietnamese characters recognition often pre-process input with a graph-based approach by trying to separate the main characters with their diacritics by determining connected regions at pixel level. This approach, however, only works well where the input contains only characters with separable diacritics, for example, scanned image of printed documents. We propose in this paper a robust method to recognize online Vietnamese characters with diacritics. Using cosine transformation with appropriated sampling algorithms, we represent multiple strokes of a character together in a single set of features. This set of features is then used as the input for a well designed machine learning based system. We have tested our system on the combination of Vietnamese characters with diacritics and Section 1c (isolated characters) of the Unipen data set, and have obtained very competitive results. © 2008 IEEE.

Index Keywords: Artificial intelligence; Information technology; Learning systems; Linguistics; Mathematical transformations; Connected regions; Data sets; Graph-based; International conferences; Language processing; Machine-learning; Pixel level; Printed documents; Process inputs; Sampling algorithms; Tonal languages; Web information; Technology

Year: 2008
Source title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Art. No.: 4584380
Page : 279-284
Link: Scopus Link
Correspondence Address: Duy, K. N.; College of Technology, Vietnam National University, Hanoi; email: s0420219@coltech.vnu.vn
Conference name: ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology
Conference date: 23 July 2008 through 25 July 2008
Conference location: Liaoning
Conference code: 73559
DOI: 10.1109/ALPIT.2008.58
Language of Original Document: English
Abbreviated Source Title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology

Document Type: Conference Paper

Source: Scopus

Authors with affiliations:

• Duy, K.N., College of Technology, Vietnam National University, Hanoi
• The, D.B., College of Technology, Vietnam National University, Hanoi

References:

• Connell, S., Jain, A., Template-based online character recognition (2001) Pattern Recognition, 34 (1)., January
• Gagne and M, C., Parizeau, Genetic engineering of hierarchical fuzzy regional representations for handwritten character recognition (2006) IJICAR, 8 (4)
• Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S., Unipen project of on-line data exchange and recognizer benchmarks (1994) Proc. of the 14th ICPR
• Li, X., Parizeau, M., Plamondon, R., Segmentation and reconstruction of on-line handwritten scripts (1998) Pattern Recognition, 31 (6)
• Mandler, E., Oed, R., Doster, W., Experiments in online script recognition (1985) Proceedings of 4th Scandinavian Conference in Image Analysis, , June
• Quan, V., Trung, P., Due, N., Ha, H., Nguyen, A., A system for recognizing vietnamese document images based on hmm and linguistics (2001) Proc. of ICDAR01
• Quan, V., Trung, P., Due, N., Ha, H., A robust method for the vietnamese handwritten and speech recognition (2002) Proc. of ICPR02