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 then 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. his 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

Abbreviated Source Title: Proceedings - ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology

Authors with affiliations:

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

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

References:

• Ahmed, N., Natarajan, T., Rao, K.R., Discrete cosine transform (1974) IEEE Trans. Computers, C-23 (1). , January

• Becker, P.W., Nielsen, K.A., Pattern recognition using dynamic pictorial information. IEEE Transaction in System, Man (1972) Cybernetics, SMC-2. , July

• Bellegarda, J.R., Nahamoo, D., Nathan, K.S., Automatic handwriting recognition using both static and dynamic parameters. (1996), United States Patent number 5,491,758, FebruaryBui, T.D., Classifying online handwriting characters under cosine representation (2007) Proc. of ALPIT 2007

• Connell, S., Jain, A., Template-based online character recognition (2001) Pattern Recognition, 34 (1). , January

• Connell, S.D., Jain, A.K., Learning prototypes for online handwritten digits (1998) Proc. of ICPR

• Gagne andM, C., Parizeau, Genetic engineering of hierarchical fuzzy regional representations for handwritten character recognition (2006) IJDAR, 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

• Hagan, M., Demuth, H., Beale, M., (1995) Neural Network Design, , PWS Publishing Company

• Hebert, J.F., Parizeau, M., Ghazzali, N., A new fuzzy geometric representation for online isolated character recognition (1998) Proc. of 14th ICPR

• Kiem, H., Bac, L., Thai, L., A fuzzy neural network for vietnamese character recognition (1999) Proceedings of ICIP99, pp. 585-589

• 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 Scandinvian Conference in Image Analysis, , June

• Parizeau, M., Lemieux, A., Gagne, C., Character recognition experiments using unipen data (2001) Proc. of 6th ICDAR

• Polyakov, V.G., Ryleev, M.A., Method and apparatus for representing image data using polynomial approximation method and iterative transformation-reparametrization technique. (1995), U.S. Patent number 5,473,742, DecemberQuan, V., Kiem, H., Trung, P., Tin, L., Ha, N., 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

• Tappert, C.C., Suen, C.Y., Wakahara, T., The state of the art in on-line handwriting recognition (1990) IEEE Transaction on Pattern Analysis and Machine Intelligence, 12 (8)

• Teredesai, A., Govindaraju, V., Issues in evolving gp based classifiers for a pattern recognition task (2004) CEC 04

• Teredesai, A., Ratzlaf, E., Subrahmonia, J., Govindaraju, V., On-line digit recognition using off-line features (2002) Proceedings of Indian Conference on Computer Vision, Graphics and Image Processing