(様式 5)

指導教員 承認印	

令和元年 12 月 11 日
Year Month Day

学位（博士）論文要旨

(Doctoral thesis abstract)

論文提出者 (Ph. D. candidate)	工学府博士後期課程　電子情報工学専攻 (major) 平成 29 年度入学(Admission year) 学籍番号　17834303　　氏名　NGUYEN CONG KHA　　印 (student ID No.)　(Name)　(Seal)
主指導教員氏名 (Name of supervisor)	中川正樹
論 文 題 目 (Title)	中国語起源の歴史文書認識の試み An Attempt to Recognize Historical Documents of Chinese Origin

論文要旨 (2000 字程度)
(Abstract(400 words))
※欧文・和文どちらでもよい。但し、和文の場合は英訳を付すこと。
(in English or in Japanese)

The recognition of Chinese original documents is interesting in many countries because they prove invaluable properties about the culture, economy, and politics of their era. To recognize historical documents, we first need a high accuracy OCR (Optical Character Recognition). In the thesis, the author presents different approaches to make OCR: over-segmentation approach, semantic segmentation based approach and segmentation-free approach. He shows the advantages and disadvantages of each approach and how to improve the recognition rate for OCR. The first approach starts with segmenting the text-line into unconnected components. Some parts of one character pattern may be segmented into small components, so it is called as over-segmented methods. Then, pre-trained OCRs are used to recognize separated components and combined components. Some unconnected components are combined together based on geometric features. The recognition result is combined with a linguistic context and a geometric context to get final results. Semantic segmentation based methods are methods that firmly segment text pages into isolated character segments, then use OCRs to recognize each of them and combine with a linguistic context to improve recognition results. The last one is the combination of a pre-trained CNN (Convolution Neural Network) and an LSTM (Long-Short Term Memory) with CTC (Connectionist Temporal Classification) without segmentation

steps. In the thesis, the author also presents a digital archiving system for the Nom script which is an ancient script used in Vietnam from the 10th to 20th century. The basic system includes four modules: pre-processing, character segmentation, character recognition and revising the result by users. From the basic system for digitizing Nom, he describes some improvements for the system with deep convolution neural networks. The improvement of the system shows better results than the previous system. Finally, this thesis presents a Character Attention Generative Adversarial Network named (CAGAN) for restoring heavily degraded character patterns in mokkan so that OCRs improve their accuracy and even help archeologists to decode them. The network is based on the U-Net like architecture with skip connections, and it is trained by the proposed loss function including the common adversarial loss (global loss) and the hierarchical character attentive loss (local loss). He made an experiment on 118 categories of most common Japanese Kanji characters, collected from severely damaged historical documents called Heijokyo mokkan written during the Nara period in Japan. The experiment shows that the method restores the shapes of characters and improves the recognition rate significantly, which is helpful for archeologists to decode damaged character patterns.