ການເພີ່ມປະສິດທິພາບການຮູ້ຈຳຕົວອັກສອນປ້າຍລິດລາວຈາກ ຮູບພາບໂດຍອັດຕະໂນມັດ ດ້ວຍວິທີ ການ jTessBoxEditor ແລະ Tesseract

ວິມົນທາ ຂຽວວົງພະຈັນ[1*], ຄອນປະເສີດ ສຸນາເຄນ[1], ສະຫວາດ ໄຊປະດິດ[1], ວິນັດ ເມກທະນາວັນ[2], ຂັນທະນຸ ຫລວງ ໄຊຊະນະ[1], ຊາລີສາ ຈັນທະວົງ[1]

ພາກວິຊາວິສະວະກຳຄອມພິວເຕີ ແລະ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ຄະນະວິສະວະກຳສາດ, ມະຫາວິທະຍາໄລແຫ່ງຊາດລາວ, ສປປ ລາວ

[*]ຕິດຕໍ່ພົວພັນ:
ວິມົນທາ ຂຽວວົງພະຈັນ, ພາກວິຊາວິສະວະກຳຄອມພິວເຕີ ແລະ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ຄະນະວິສະວະກຳສາດ, ມະຫາວິທະຍາໄລແຫ່ງຊາດລາວ, ເບີໂທ: +856 20 5559 2414,
ອີເມວ:
vimontha@fe-nuol.edu.la
[1]ພາກວິຊາວິສະວະກຳຄອມພິວເຕີ ແລະ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ຄະນະວິສະວະກຳສາດ, ມະຫາວິທະຍາໄລແຫ່ງຊາດລາວ,
[2]ຄະນະວິສະວະກຳສາດ, ມະຫາວິທະຍາໄລສຸພານຸວົງ

ຂໍ້ມູນບົດຄວາມ:
ການສົ່ງບົດ: 05 ກັນຍາ 2023
ການປັບປຸງ: 20 ພະຈິກ 2023
ການຕອບຮັບ: 05 ທັນວາ 2023

ບົດຄັດຫຍໍ້

ການຮູ້ຈຳຕົວອັກສອນ ດ້ວຍແສງ OCR ເປັນເຕັກໂນໂລຊີ ສຳຄັນທີ່ແປງ ຂໍ້ຄວາມທີ່ພິມ ຫຼື ຂຽນດ້ວຍລາຍມືໃຫ້ເປັນຂໍ້ຄວາມທີ່ເຄື່ອງສາມາດອ່ານໄດ້ ເຮັດ ໃຫ້ສາມາດປະມວນຜົນ ແລະ ວິເຄາະຂໍ້ມູນໄດ້ໂດຍອັດຕະໂນມັດແຕ່ເຖິງຢ່າງໃດກໍ່ ຕາມ OCR ສຳລັບ ພາສາທີ່ມີຂໍ້ຄວາມຊັບຊ້ອນ ເຊັ່ນພາສາລາວແມ່ນສິ່ງທີ່ທ້າທາຍ ເນື່ອງຈາກວ່າລັກສະນະຕົວອັກສອນທີ່ຊັບຊ້ອນ. ດັ່ງນັ້ນ, ບົດຄົ້ນຄວ້າວິໄຈໄດ້ນຳສະ ໜີ ການໃຊ້ເຄື່ອງມື jTessBoxEditor ສຳລັບການຝຶກ OCR ໃຫ້ຮູ້ຈຳຂໍ້ຄວາມ ພາສາລາວ. ເຄື່ອງມື jTessBoxEditor ແມ່ນສ່ວນເສີມຂອງ ເຄື່ອງມື Tesseract OCR ມີສ່ວນໜ້າຕ່າງທີ່ໃຊ້ງານງ່າຍ ສຳລັບການສ້າງ ແລະ ປັບແຕ່ງຂໍ້ມູນການຝຶກ. ບົດຄົ້ນຄວ້າວິໄຈໄດ້ສະແດງຂະບວນການສ້າງແບບຈຳລອງ OCR ສຳລັບຂໍ້ຄວາມ ພາສາລາວ ລວມເຖິງການເກັບຂໍ້ມູນ, ການສ້າງການຝຶກຂໍ້ມູນ ແລະ ການວັດປະເມີນ ຜົນການສ້າງແບບຈຳລອງ. ຜົນການທົດເຫັນວ່າ ການສ້າງຊຸດຂໍ້ມູນ,ການສ້າງໄຟລ Tiff/Box ໂດຍການແຍກຕົວອັນສອນ, ແຍກຕົວພະຍັນຊະນະ ຫຼື ການປະສົມ ພະຍັນຊະນະ ແລະ ຕົວອັກສອນ ການຮູ້ຈຳຕົວອັກສອນພາສາລາວໃນການປ້າຍລິດ ລາວແມ່ນມີປະສິດ ທິພາບຄືກັນ ໂດຍມີການຝຶກແບບ train with existing box ແລະ ການຝຶກແບບ shape clustering. ໃນສ່ວນຄວາມໄວ ແລະ ຄວາມບໍ່ຊັບຊ້ອນ ເຫັນວ່າ ການສ້າງຊຸດຂໍ້ມູນແບບການແຍກຕົວອັກສອນ, ແຍກພະຍັນຊະນະ ແລະໃຊ້ ການຝຶກແບບ train with existing box ແມ່ນມີຜົນດີກວ່າ.

ຄຳສັບສຳຄັນ: *ການຮູ້ຈຳຕົວອັກສອນ, ຕົວອັກສອນລາວ, ຊຸດຂໍ້ມູນ, ການຝຶກ, ປ້າຍ ລິດລາວ, ເຄື່ອງມື jTessBoxEditor, ການປະມວນຜົນຮູບພາບ*

Enhancing Lao Vehicle License Plate Recognition through jTessBoxEditor and Tesseract Optical Character Recognition (OCR)

Vimontha Khieovongphachanh[1*], Khonepaserth Sounakhen[1], Savath Saypadith[1], Vinath Mekthanavanh[2], Khanthanou Luangxaysana[1], Xalisa Chanthavong[1]

Department of Computer Engineering and Information Technology, Faculty of Engineering, National University of Laos, Lao PDR

Correspodence:
Vimontha Khieovongphachanh, Department of Computer Engineering and Information Technology, Faculty of Engineering, National University of Laos, PDR
Tel: +856 20 5559 2414,
Email: vimontha@fe-nuol.edu.la
[1]Department of Computer Engineering and Information Technology, Faculty of Engineering, National University of Laos, Lao PDR
[2]Faculty of Engineering, Souphanouvong University

Article Info:
Submitted: Sep 05, 2023
Revised: Nov 20, 2023
Accepted: Dec 05, 2023

Abstract

Optical Character Recognition (OCR) is a critical technology that converts printed or handwritten text into machine-readable text, enabling automated data processing and analysis. However, OCR for languages with complex text, such as Lao, presents unique challenges due to the intricate nature of the characters. This paper introduces the use of the jTessBoxEditor tool for training an OCR engine to recognize Lao text. The jTessBoxEditor tool, an extension of the Tesseract OCR engine, provides a user-friendly interface for creating and refining training data. The experimental results demonstrate the effectiveness of various techniques in recognizing Lao characters on Lao car plate. These techniques include data set creation, the generation of Tiff/Box files through character separation, and the isolation or combination of consonants and letters. Notably, the utilization of both existing box training and shape clustering training contributes to improved recognition performance. Furthermore, our findings highlight the importance of speed and simplicity in the OCR modeling process. Specifically, the creation of a dataset that involves character separation and consonant isolation, coupled with the use of existing box training, emerges as the most efficient and effective approach for Lao text recognition.

Keyword: *Optical Character Recognition (OCR), Lao Character, Data Set, Data training, Lao Car Plate, jTessBoxEditor tool, Image Processing.*

1. ພາກສະເໜີ

ເຕັກໂນໂລຊີ OCR (Optical Character Recognition) ແມ່ນການແຍກຂໍ້ຄວາມອອກຈາກຮູບພາບ ຫຼື ການແປງຂໍ້ຄວາມທີ່ຢູ່ໃນຮູບພາບມາເປັນຂໍ້ຄວາມດິຈິຕ້ອນໂດຍການສາຍແສງ ເຊິ່ງມີການລິເລີ່ມຕັ້ງແຕ່ປີ ຄ.ສ 1920 ໂດຍນັກຟິຊິກຊາວເຢຍລະມັນທີ່ມີຊື່ວ່າ Emanuel Goldberg ໄດ້ປະດິດອຸປະກອນທີ່ເປັນຈຸດເລີ່ມຕົ້ນຂອງເຕັກໂນໂລຊີ OCR ເຊິ່ງເປັນຕົວອ່ານຕົວອັກສອນແລ້ວປ່ຽນເປັນລະຫັດໂທລະເລກມາດຕະຖານ (Phommachanh & Phommachanh, 2019). ຫຼັງຈາກນັ້ນໃນປີ ຄ.ສ 1927 Goldberg ກໍໄດ້ພັດທະນາເຕັກໂນໂລຊີ OCR ໃຫ້ມີຄຸນນະພາບດີກວ່າເກົ່າກັບດ້ວຍສິ່ງປະດິດທີ່ມີຊື່ວ່າ “Statistical Machine” ເຄື່ອງມືທີ່ເຮັດໜ້າທີ່ຊອກຫາບ່ອນເກັບມ້ຽນໄມໂຄຣຟິມ (Micro film) ເຊິ່ງເຮັດໜ້າທີ່ ໃນການລະບຸຂໍ້ມູນດ້ວຍ ການຈີ່ດ້ວຍແສງ ໂດຍການເຮັດວຽກຂອງເຄື່ອງອ່ານສະຖິຕິໄດ້ເຮັດໃຫ້ Goldberg ປະສົບຜົນສຳເລັດ ແລະ ໄດ້ຮັບສິດທິບັດສຳລັບການປະດິດນີ້ກ່ອນທີ່ຈະຖືກຊື້ໂດຍ IBM. ຈົນມາເຖິງປີ ຄ.ສ 2000 ທີ່ OCR ອອນໄລນ໌ (Web OCR) ຖືກເປີດໃຊ້ຜ່ານຄລາວ (Cloud) ແລະ ແອັບພລິເຄຊັນເທິງມືຖື (Mobile Application) ທີ່ເພີ່ມຄວາມສະດວກໃນການນຳໃຊ້ໂດຍຜ່ານໂທລະສັບມືຖືທີ່ມີຄວາມສະຫຼາດຫຼາຍຂຶ້ນ ເຊິ່ງເຮັດໜ້າທີ່ໂດຍການແຍກຂໍ້ ຄວາມທີ່ບັນທຶກໄວ້ໂດຍໃຊ້ກ້ອງຖ່າຍຮູບຂອງອຸປະກອນ ຖ້າອຸປະກອນບໍ່ຮອງຮັບ OCR ກໍຍັງມີວິທີການອ່ານຕົວອັກສອນໂດຍໃຊ້ OCR API ທີ່ເຮັດວຽກໂດຍການແຍກຂໍ້ຄວາມອອກຈາກໄຟລ໌

ຮູບພາບ, ຈາກນັ້ນ ມັນຈະສົ່ງຄືນຂໍ້ຄວາມທີ່ແຍກອອກມາ ໄປຫາແອັບພລິເຄຊັນອຸປະກອນເພື່ອດຳເນີນການຕໍ່ໄປ.

ເຕັກໂນໂລຊີ OCR (Kaur & Singh, 2019; Kumar & Singh, 2022) ມີຄວາມສຳຄັນຫຼາຍ ແລະ ຍັງ ໄດ້ຖືກນຳໃຊ້ຢ່າງຫຼວງຫຼາຍໃນຊີວິດປະຈຳວັນຂອງພວກ ເຮົາຍົກຕົວຢ່າງ: ເຄື່ອງສະແກນເອກະສານ, ແອັບແປພາສາ, ເຄື່ອງມືປ່ຽນຂໍ້ຄວາມຈາກຮູບພາບເປັນຂໍ້ຄວາມຕົວອັກ ສອນ ແລະ ອື່ນໆ ເຊິ່ງໃນເຕັກໂນໂລຊີ OCR ຫຼື ການຮູ້ຈຳ ອັກສອນພາສາລາວມີຄວາມທ້າທາຍ ເນື່ອງຈາກພາສາລາວ ມີເອກະລັກ ແລະ ໂຄງສ້າງທີ່ຊັບຊ້ອນ ບໍ່ຄືກັບຕົວອັກສອນ ພາສາອັງກິດ ເຊິ່ງການຄົ້ນຄວ້າວິໃຈ ແລະ ການນຳໃຊ້ ແມ່ນຍັງມີຄວາມຖືກຕ້ອງຂອງຂໍ້ມູນຍັງຕ່ຳ ແລະ ຍັງບໍ່ຖືກ ໄວຍະກອນໃນພາສາລາວ ເຮັດໃຫ້ຫຼາຍພາກສ່ວນຍັງບໍ່ສາ ມາດໃຊ້ເຄື່ອງມືທີ່ອຳນວຍຄວາມສະດວກໃນການສ້າງ ຫຼື ແປເອກະສານ. ການຄົ້ນຄວ້າທີ່ຜ່ານມາ ສຳລັບພາສາທີ່ຊັບ ຊ້ອນ ເຊັ່ນ: ພາສາອາລາບິກ ຫຼື ພາສາຈີນ ເຊິ່ງຂໍ້ມູນການ ຝຶກທີ່ມີປະສິດ ທິພາບສູງ ແລະ ເຄື່ອງມື jTessBoxEditor (Singh, & Singh, 2018; Sharma, & Gupta, 2019) ໄດ້ຮັບໃນການສ້າງແບບຈຳລອງການຈື່ຈຳຕົວອັກສອນທີ່ ຊັບຊ້ອນປະກອບກັບການຮູ້ຈຳອັກສອນພາສາລາວຍັງບໍ່ທັນ ມີການເຜີຍ ແຜ່. ດັ່ງນັ້ນ, ບົດຄົ້ນຄວ້າວິໃຈນີ້ ຈຶ່ງມີຈຸດປະ ສົງນຳໃຊ້ການຝຶກການຮູ້ຈຳຕົວອັກສອນລາວ ດ້ວຍ jTessBoxEditor ແລະ Serack Tesseract Trainer (Ammar & Koubaa, & Boulila & Benjdira & Alhabashi, 2023), (Yap, W. J., & Tan, C. C. ,2019), (Hlaing, Z. T., Zaw, Z., & Oo, K. Z. ,2018) ເພື່ອສ້າງແບບຈຳລອງ ການຮູ້ຈຳພາສາລາວ ທີ່ ຖືກຕ້ອງສຳລັບຕົວອັກສອນ ໂດຍ ປະຍຸກໃຊ້ກັບ ປ້າຍລົດ ໃນປະເທດລາວ.

2. ອຸປະກອນ ແລະ ວິທີການ

2.1 ເຄື່ອງມື jTessBoxEditor

jTessBoxEditor ເປັນເຄື່ອງມືຕົວຊ່ວຍທີ່ມີໜ້າ ຕ່າງກຣາຟຟິກສຳລັບຜູ້ໃຊ້ (Graphical User Interface: GUI) ເຊິ່ງເປັນເຄື່ອງມືຊ່ວຍໃນການສ້າງ, ແກ້ໄຂ ແລະ ປັບ ປຸງຂໍ້ມູນການຝຶກສຳລັບ Tesseract OCR (Khan & Khan, 2019) ໂດຍ Tesseract ເປັນເຄື່ອງມື OCR ແບບ Open Source ອີກຕົວນຶ່ງທີ່ມີຄວາມສາມາດໃນການຮູ້ຈຳ ພື້ນຖານ ໃນຂະນະທີ່ jTessBoxEditor ເປັນເຄື່ອງມືຕົວ ຊ່ວຍທີ່ມີໜ້າຕ່າງກຣາຟຟິກສຳລັບຜູ້ໃຊ້ (Graphical User Interface: GUI) ແບບ Open source ທີ່ໃຊ້ສຳລັບແກ້ ໄຂ້ໄຟລ໌ຂໍ້ມູນການ ການຝຶກ ທີ່ໃຊ້ໃນລະບົບການຮູ້ຈຳ ໄດ້ ຮັບການອອກແບບມາໂດຍສະເພາະສຳລັບການເຮັດວຽກ ຮ່ວມກັບ Tesseract OCR ເຊິ່ງເປັນເຄື່ອງມື OCR ທີ່ ພັດທະນາໂດຍ Google jTessBoxEditor ຊ່ວຍໃຫ້ຜູ້ໃຊ້ ສາມາດແກ້ໄຂໄຟລ໌ບ່ອກ (Box) ເຊິ່ງມີຂໍ້ມູນກ່ຽວກັບຕຳແ ໜ່ງ ແລະ ຂະໜາດຂອງລັກສະນະ ຫຼື ອັກສອນແຕ່ລະຕົວໃນ ຮູບພາບໄຟລ໌ບ່ອກເຫຼົ່ານີ້ຈຳເປັນສຳລັບການຝຶກ Tesseract OCR ໃຫ້ຮູ້ຈຳຂໍ້ຄວາມໃນຮູບພາບໄດ້ຢ່າງຊັດເຈນ ດ້ວຍ jTessBoxEditor ຜູ້ໃຊ້ສາມາດເປີດໄຟລ໌ບ່ອກ (Box) ສະແດງພາບທີ່ກ່ຽວຂ້ອງ ແລະ ທຳການປັບປ່ຽນ Box ດ້ວຍ ຕົນເອງຖ້າຈຳເປັນ. Tesseract ຖືກພັດທະນາໂດຍໂປຣ ແກຣມເມີ ທີ່ມີຊື່ວ່າ Serak Shiferaw ເຊິ່ງໂປຣແກຣມ ເປັນໂປຣແກຣມສຳລັບ ຝຶກ OCR Model (Phommachanh & Phommachanh, 2020) ແບບ ສຳເລັດຮູບຈາກໄຟລ໌ Tiff/Box ພ້ອມທັງກຳນົດອັກສອນ ສຳລັບ Train ໄດ້ນຳອີກ ເຊິ່ງການ Train Model ສຳເລັດ ຈະໄດ້ ໄຟລ໌ Model ສຳລັບໃຊ້ກັບ Tesseract OCR. ນຳ ໃຊ້ງານ Serak Tesseract trainer ໃນການນຳໄຟລ໌ Tiff/Box ມາ Train ເປັນ OCR Model ແລະ ຕັ້ງຄ່າຕົວ ອັກສອນຕ່າງໆ ຂອງ Model. ເຊິ່ງຖືກນຳໃຊ້ໃນຕົວສອນ ລາຕິນ.

2.2 ວິທີການ

ແມ່ນການສຸ່ມປ້າຍລົດລາວ 100 ປ້າຍ ເພື່ອທຳການ ອ່ານໃຫ້ເປັນຕົວເລກ ແລະ ຕົວອັກສອນ ແລະ ທົດສອບ ຄວາມແມ່ນຍຳຄວາມຖືກຕ້ອງຂອງອ່ານປ້າຍລົດລາວ.

ຂັ້ນຕອນການທົດລອງ ດັ່ງ ຮູບພາບທີ 2 ຂັ້ນຕອນ ການຈົດຈຳຕົວເລກ ແລະ ຕົວອັກສອນ ປ້າຍລົດລາວ.

2.2.1 ກຽມຊຸດຂໍ້ມູນ

- ຕົວພະຍັນຊະນະລາວປະກອບມີທັງໝົດ: 27 ໂຕ
- ຕົວສະຫຼະລາວປະກອບມີທັງໝົດ: 27 ໂຕ
- ຕົວເລກປະກອບມີທັງໝົດ: 10 ໂຕ

2.2.2 ຊຸດຂໍ້ມູນ

ຊຸດຂໍ້ມູນສຳລັບໃຫ້ Machine Learning (Kumar & Kumar, 2018; Kaur & Singh, 2019) ຮຽນຮູ້ແມ່ນຂໍ້ມູນທີ່ຈະນຳເຂົ້າລະບົບເພື່ອເຮັດໃຫ້ Machine Learning ຮຽນຮູ້ໂດຍລັກສະນະຂອງຂໍ້ມູນຈະ ປະກອບດ້ວຍຄຸນລັກສະນະ (Features) ແລະ ຄຳຕອບ ຂອງການຮຽນຮູ້ຊຸດຂໍ້ມູນນັ້ນ (Label).

- Label ແມ່ນຂໍ້ມູນທີ່ເຮົາຕ້ອງການຈະຄາດເດົາ (Predict) ໃນ Linear Regression Model.

- Features ແມ່ນຄຸນລັກສະນະທີ່ ມີລັກສະນະສະເພາະຂອງຂໍ້ມູນທີ່ເຮົາຕ້ອງການຈະຮຽນຮູ້

ໂດຍເຮົາຈະໃຊ້ຂໍ້ມູນເຫຼົ່ານີ້ໃນການ ເຝິກ ແລະ ການຄາດເດົາ.

2.2.3 ຂັ້ນຕອນການຝຶກ

ນຳໃຊ້ jtessboxeditor ສ້າງໄຟລ໌ Tiff/Box ຈາກຊຸດຂໍ້ມູນ ເຊິ່ງໄຟລ໌ນີ້ຈະເປັນການຕັດຄຳ.

ຫຼັງຈາກໄດ້ໄຟລ໌ Tiff/Box ແລ້ວກໍນຳມາຝຶກ ໃນໂປຣແກຣມ Serak Trainer ເພື່ອສ້າງ Model OCR ທີ່ສາມາດໃຊ້ໄດ້ກັບຫຼາຍໆ Tools

ວິທີການສ້າງໄຟລ Tiff/Box ຈະແບ່ງເປັນ 2 ຮູບແບບ:

- ຮູບແບບທີ1(Tiff/Box1): ການສ້າງ Tiff/Box ໂດຍແຍກແຕ່ລະຕົວພະຍັນຊະນະ ຫຼື ແຕ່ລະຕົວສະຫຼະ
- ຮູບແບບທີ2(Tiff/Box2): ການສ້າງ Tiff/Box ໂດຍການປະສົມພະຍັນຊະນະ ແລະ ຕົວສະຫຼະການຝຶກ Tiff/Box ຈະມີແບບ 2 ແບບ:
  - ການຝຶກແບບທີ 1: train with existing box
  - ການຝຶກແບບທີ 2: shape clustering

2.2.4 ການທົດລອງ

ສຳລັບຂັ້ນຕອນການທົດລອງແມ່ນເຮົາຈະນຳຂໍ້ມູນຮູບປ້າຍລົດ (ຈຳລອງ) ເຊິ່ງປ້າຍລົດທີ່ເປັນຮູບພາບຈະມີຂໍ້ມູນທີ່ເປັນ Text ໄຟລ໌ເປັນໄຟລ໌ .txt ເປັນຂໍ້ມູນປ້າຍທີ່ເປັນຕົວອັກສອນຕົ້ນສະບັບ ແລະ ຈະມີຂໍ້ມູນປ້າຍລົດທີ່ເປັນໄຟລ໌ PNG ເຊິ່ງເຮົາຈະນຳເອົາໄຟລ໌ ຂໍ້ມູນທີ່ເປັນໄຟລ໌ PNG ເຂົ້າຂະບວນການ OCR (Optical Character Recognition) ເປັນເຊິ່ງຈະນຳເອົາຂໍ້ມູນທີ່ເປັນຮູບພາບມາເຂົ້າຂະບວນການ OCR (Optical Character Recognition) ເປັນຂະບວນການທີ່ມີການຈັດຊຸດຂໍ້ມູນແບ່ງອອກເປັນ 2 ແບບ: Tiff/ Box1 (ຂໍ້ມູນທີ່ມີຮູບແບບຮູບ ແບບເປັນການແຍກສະຫະ, ພະຍັນຊະນະ ແລະ ຕົວເລກ) ແລະ Tiff/Box2 (ຂໍ້ມູນທີ່ມການປະສົມພະຍັນຊະນະ ແລະ ສະຫະ) ແລ້ວນຳຝຶກ. ໂດຍການຝຶກຈະປະກອບມີ 2 ຮູບແບບຄື: Train with existing box ແລະ Shape Clustering ເຊິ່ງເມື່ອຜ່ານຂະບວນການ OCR (Optical Character Recognition) ເຮົາກໍຈະໄດ້ຮັບຂໍ້ມູນອັກສອນທີ່ຢູ່ໃນຮູບພາບໂດຍຈະເປັນອັກສອນດິຈິຕອນທີ່ຜ່ານຂະບວນການມາແລ້ວ ແລະ ນຳເອົາຕົວອັກສອນທີ່ຜ່ານການຝຶກທັງ 2 ຮູບແບບມາປຽບທຽບກັບອັກສອນຕົ້ນສະບັບ.

2.2.5 ຜົນການທົດລອງ

ຜົນການທົດລອງຂອງການຝຶກຮູບແບບ Train with existing box ແລະ Shape Clustering ແລະ ນຳໃຊ້ຂໍ້ມູນທັງສອງຊຸດເຊິ່ງຂໍ້ມູນທີ່ນຳມາຝຶກປະກອບມີ: ພະຍັນຊະນະ, ສະຫຼະ ແລະ ຕົວເລກ ຄື Tiff/Box1 (ຂໍ້ມູນທີ່ມີຮູບແບບເປັນການແຍກສະຫຼະ, ພະຍັນຊະນະ ແລະ ຕົວເລກ) ແລະ Tiff/Box2 (ຂໍ້ມູນທີ່ມີການປະສົມພະຍັນຊະນະ ແລະ ສະຫຼະ) ຜ່ານຂະບວນການ OCR ອີງຕາມຕາຕະລາງທີ 4 ສັງເກດໄດ້ຜ່ານຕາຕະລາງຈະປະກອບມີ ຕົວອັກສອນຕົ້ນແບບ 20 ຊຸດທີ່ເຮົານຳມາໃຊ້ໃນການຝຶກ ແລະ ຜົນລັບຂອງການຝຶກທັງສອງແບບໂດຍນຳໃຊ້ຊຸດຂໍ້ມູນທັງສອງແບບ Tiff/Box1 ແລະ Tiff/Box2 ລວມມີຜົນລັບທັງໝົດ 4 ແບບ: Train with existing box (Tiff/Box1), Train with existing box (Tiff/Box2), Shape Clustering (Tiff/Box1), Shape Clustering(Tiff/Box2) ເຊິ່ງແຕ່ລະແບບຈະໃຫ້ຜົນຮັບທີ່ແຕກຕ່າງກັນ. ສຳລັບຂໍ້ຄວາມທີ່ຜ່ານການຝຶກມາແລ້ວແຕ່ຜົນຮັບທີ່ຍັງບໍ່ເປັນທີ່ພຶງພໍໃຈແມ່ນມີ 5 ຊຸດຂໍ້ມູນ: ຫຼວງນ້ຳທາ (2), ຜົ້ງສາລີ (2), ສາລະວັນ ເນື່ອງຈາກວ່າຂໍ້ຄວາມດັ່ງກ່າວມີຄວາມຊັບຊ້ອນຈຶ່ງເຮັດຜົນລັບທີ່ອອກມາບໍ່ດີປານໃດ. ສ່ວນຊຸດຂໍ້ມູນອື່ນໆແມ່ນກໍໄດ້ຜົນລັບທີ່ໃກ້ຄຽງກັບຊຸດຂໍ້ມູນຕົ້ນແບບ ແຕ່ກໍຍັງມີບາງຕົວອັກສອນ, ສະຫຼະ, ພະຍັນຊະນະ ທີ່ຍັງສະກົດຜິດຢູ່ບາງສ່ວນ. ສຳລັບຕົວເລກແມ່ນໄດ້ຮັບຜົນລັບທີ່ດີຫຼາຍຊຸດຂໍ້ມູນແຕ່ລະຊຸດທີ່ຜ່ານການຝຶກແລ້ວມີຄວາມແນ່ນອນຫຼາຍທີ່ສຸດລ້ວນແຕ່ແມ່ນຂໍ້ຄວາມທີ່ເປັນຕົວເລກ.

2.2.6 ການທົດສອບຄວາມແມ່ນຢຳ

ການຄິດໄລ່ຄວາມແມ້ນຢຳ ແມ່ນໃຊ້ວິທີ Levenshtein Distance Equation (Nam, 2018)

$$lev_{a,b}(i,j) = \{max(i,j) if min(i,j) = 0 min\{lev_{a,b}(i-1,j) + 1 lev_{a,b}(i,j-1) + 1 lev_{a,b}(i-1,j-1) + 1_{(\alpha_i \neq b_j)} otherwise \quad (1)$$

$$accuracyrate = (1 - \frac{lev}{(length(a),length(b))}) * 100 \quad (2)$$

a: ຕົວອັກສອນຕົ້ນສະບັບ

b: ຕົວອັກສອນຈາກຂະບວນການ OCR

i: ຕຳແໜ່ງຕົວອັກສອນຕົ້ນສະບັບ

j: ຕຳແໜ່ງຕົວອັກສອນຈາກຂະບວນການ OCR

- Levenshtein Distance (lev): ຈຳນວນຄວາມແຕກຕ່າງລະຫວ່າງ ຕົວອັກສອນຕົ້ນສະບັບ ແລະ ຕົວອັກສອນຈາກຂະບວນການ OCR.

- accuracyrate: ຄວາມແມ່ນຍຳລະຫວ່າງ ຕົວອັກສອນຕົ້ນສະບັບ ແລະ ຕົວອັກສອນຈາກຂະບວນການ OCR ຄິດໄລ່ ເປັນເປີເຊັນ.

3. ຜົນໄດ້ຮັບ

ຜົນຂອງການເຝິກແບບທີ 1 ຈະເຫັນໄດ້ວ່າ Tiff/Box ຮູບແບບທີ 1 ຈະມີຄ່າຄວາມແມ່ນຍຳສູງກວ່າ Tiff/Box ຮູບແບບທີ 2 ທຸກປ້າຍລິດທີ່ນຳມາທົດລອງ ແລະ ມີຜົນຄ່າສະເລ່ຍຄວາມແມ່ນຍຳເປັນເປີເຊັນ 60.32% ໃນຂະນະທີ່ການເຝິກ Tiff/Box ຮູບແບບທີ 2 ເຫັນວ່າ ຄ່າຄວາມແມ່ນຍຳຂອງ Tiff/Box 1 ແລະ Tiff/Box 2 ມີຄ່າໃກ້ຄຽງກັນທຸກປ້າຍລິດ ແລະ ມີຜົນຄ່າສະເລ່ຍຄວາມແມ່ນຍຳເປັນເປີເຊັນ 59.11% ເຊິ່ງເຫັນໄດ້ວ່າຄວາມແມ່ນຍຳຂອງຮູບແບບທີ 1 ຈະສູງກວ່າ ແລະ ບໍ່ມີຂັ້ນຕອນໃນການເຝິກຂໍ້ມູນທີ່ຊັບຊ້ອນ ຖ້າການທົດລອງມີການເຝິກຂໍ້ມູນຫຼາຍຂຶ້ນກໍ່ຈະເຮັດໃຫ້ຜົນຂອງຮູບແບບທີ 1 ມີຄວາມແມ່ນຍຳສູງຂຶ້ນ.

- ການຝຶກແບບທີ 1: train with existing box ໃນTiff/Box 2 ຮູບແບບ. ດັ່ງຕາຕະລາງທີ 2

- ການຝຶກແບບທີ 2: Shape Clustering ໃນ Tiff/Box 2 ຮູບແບບ. ດັ່ງຕາຕະລາງທີ 3

4. ວິພາກຜົນ

ການຄົ້ນຄວ້າກ່ຽວກັບ OCR ຂອງພາສາລາວກໍ່ຄືການຈຳຕົວອັກສອນໃນຮູບແບບຂໍ້ຄວາມແມ່ນຍັງບໍ່ແຜ່ຫຼາຍ ແລະ ເຕັກນິກການເຝິກ Machine Learning ມີຮູບແບບເຕັກນິກທີ່ແຕກຕ່າງກັນເຊິ່ງການນຳໃຊ້ຂຶ້ນກັບຈຸດປະສົງທີ່ແຕກຕ່າງກັນ ແລະ ເຫັນວ່າແມ່ນມີຄວາມແມ່ນຍຳ 90% ຂຶ້ນໄປ ເຊິ່ງເຫັນວ່າມີຄວາມແມ່ນຍຳສູງ ຍ້ອນຂໍ້ມູນມີຂະໜາດໃຫຍ່ໃນການເຝິກ ແລະ ເຕັກນິກມີຄວາມຊັບຊ້ອນ ເຮັດໃຫ້ມີການປະມວນຜົນໃຊ້ເວລາຫຼາຍຂຶ້ນ (Kesom et al, 2018) (Beyene et al, 2019). ບົດຄົ້ນຄວ້ານີ້ສຶກສາເຕັກນິກທີ່ບໍ່ມີຄວາມຊັບຊ້ອນ ແລະ ການເຝິກຂໍ້ມູນບໍ່ຫຼາຍສຳລັບພາສາລາວ ເຊິ່ງເຕັກນິກທີ່ນຳມາໃຊ້ນີ້ແມ່ນຖືກນຳໃຊ້ກັບພາສາຕ່າງປະເທດ ເຊັ່ນ: ພາສາລາຕິນ ແລະ ພາສາພື້ນເມືອງອື່ນໆ ເຫັນວ່າ Tesseract OCR ມີຄວາມແມ່ນຍຳໃນລະດັບທີ່ໜ້າພໍໃຈ, ເຊິ່ງມີຄວາມແຕກຕ່າງກັບບົດຄົ້ນຄວ້າການນຳໃຊ້ເຄື່ອງມືໃຊ້ໃນການອ່ານເອກະສານພາສາລາຕິນ (Khadijaelgajoui et al., 2015) ແມ່ນສາມາດປັບປຸງການອ່ານໄດ້ດີ ແຕ່ຍັງບໍ່ທັນໄດ້ລົງເລິກ ແລະ ຍັງບໍ່ທັນເພີ່ມ ລັກສະນະຂອງຟອນພາສາ. ບົດຄົ້ນຄວ້າອ່ານຂໍ້ຄວາມຈາກເອກະສານ PDF ໃນພາສາທີ່ມີຊັບພະຍາກອນຕ່ຳໂດຍໃຊ້ຕົວອັກສອນພື້ນເມືອງ, ບົດຄົ້ນ ຄວ້າໄດ້ປັບປຸງການປະຕິບັດຂອງ Tesseract, ເຄື່ອງມືການຮັບຮູ້ຕົວອັກສອນໂດຍການຝຶກມັນຢູ່ໃນຫຼາຍກວ່າ 20 ຕົວອັກສອນພື້ນເມືອງໃນ Tamil ແລະ Sinhala (Charangan Vasantharajan et al., 2019). ວິທີການສຳລັບການສະກັດຂໍ້ຄວາມຈາກເອກະສານ PDF ໂດຍໃຊ້ຕົວອັກສອນພື້ນເມືອງ. ຜູ້ຂຽນໄດ້ປັບປຸງການປະຕິບັດຂອງ Tesseract, ເຄື່ອງມືການຮັບຮູ້ຕົວອັກສອນ optical (OCR), ໂດຍການຝຶກອົບຮົມມັນຢູ່ໃນຫຼາຍກວ່າ 20 ຕົວອັກສອນພື້ນເມືອງໃນ Tamil ແລະ Sinhala. ວິທີກາ ນນີ້ຊ່ວຍຫຼຸດອັດຕາຄວາມຜິດພາດໃນລະດັບຕົວອັກສອນຂອງ Tesseract ແລະ ສະແດງໃຫ້ເຫັນວ່າ ການໃຊ້ຫຼັກການສຳລັບອັກສອນພື້ນເມືອງໃນ Tamil ແລະ Sinhala ແມ່ນສາທດເຂົ້າໃຈຂໍ້ຄວາມ ແລະ ປັບປຸງຂະບວນການ OCR ແລະ ມີວິທີການຝຶກແບບໃໝ່. ເຄື່ອງອ່ານຂໍ້ຄວາມທີ່ຂຽນດ້ວຍມືຊ່ວຍຄົນພິການທາງສາຍຕາໂດຍການປ່ຽນຂໍ້ຄວາມທີ່ສະແກນ (hithra Selvaraj & Bhalaji Natarajan, 2018), ພິມ ແລະ ຂຽນດ້ວຍມືເປັນສຽງ. ເຄື່ອງສະແກນໜ້າມືຖືຖືກນຳໃຊ້ເພື່ອສະແກນຂໍ້ຄວາມ. ແອັບພລິເຄຊັນນີ້ໃຊ້ເຄື່ອງມື Tesseract OCR ເພື່ອອ່ານຂໍ້ຄວາມຈາກຮູບພາບ ແລະ ປ່ຽນເປັນສຽງເວົ້າ. ເຄື່ອງມື OCR ໄດ້ຖືກຝຶກໃຫ້ຮັບຮູ້ຂໍ້ຄວາມທີ່ຂຽນດ້ວຍມືສຳລັບ ພາສາອິນດູ ແລະ Bengali.

ຜ່ານການຝຶກທັງສອງຮູບແບບ ທັງ Train with existing box ແລະ Shape Clustering ສັງເກດໄດ້ວ່າ ຜົນຮັບທີ່ທີ່ໄດ້ຮັບຄວາມແມ່ນຍຳດີທີ່ສຸດຈາກຮູບແບບ Train with existing box ແມ່ນການນຳໃຊ້ຊຸດຂໍ້ ມູນ Tiff/Box1 (ຂໍ້ມູນທີ່ມີຮູບແບບເປັນການແຍກສະຫຼະ, ພະຍັນຊະນະ ແລະ ຕົວເລກ), ສ່ວນຜົນລັບທີ່ໄດ້ຮັບຄວາມແມ່ນຍຳທີ່ດີທີ່ສຸດຂອງຮູບແບບ Shape Clustering ແມ່ນການນຳໃຊ້ຊຸດຂໍ້ ມູນ Tiff/Box2 (ຂໍ້ ມູນທີ່ມີການປະສົມພະຍັນຊະນະ ແລະ ສະຫຼະ). ຈາກຜົນລັບການປຽບທຽມຄວາມແມ່ນຍຳ ຈາກຮູບພາບທີ 10 ສັງເກດໄດ້ວ່າ ຮູບແບບ Train with existing box ທີ່ນຳໃຊ້ຊຸດຂໍ ມູນ Tiff/Box1 (ຂໍ້ ມູນທີ່ມີຮູບແບບເປັນການແຍກສະຫຼະ, ພະຍັນຊະນະ ແລະ ຕົວເລກ) ແມ່ນມີຄວາມແມ່ນຍຳຫຼາຍກວ່າ ຮູບແບບ Shape Clustering ທີ່ນຳໃຊ້ຊຸດຂໍ ມູນ Tiff/Box2 (ຂໍ ມູນທີ່ ມີການປະສົມພະຍັນຊະນະ ແລະ

ສະຫຼະ) ບໍ່ວ່າຈະເປັນເປີເຊັນຄວາມແມ່ນຍຳ ຫຼື ຜົນຮັບຜ່ານ ການຝຶກແມ່ນ ຮູບແບບ Train with existing box ທີ່ນຳໃຊ້ຊຸດຂໍ້ມູນ Tiff/Box1 ເຮັດໄດ້ດີກວ່າຫຼາຍ ເຖິງຈະໄດ້ຮັບຊຸດຂໍ້ມູນທີ່ຊັບຊ້ອນແຕ່ກໍໄດ້ຜົນຮັບທີ່ດີ ແລະ ໃກ້ຄຽງກັບຊຸດຂໍ້ມູນທີ່ນຳມາຝຶກ. ຕົວຢ່າງ: ຕົວອັກສອນຕົ້ນສະບັບ (ເຊກອງ, ອັດຕະປື, ຫົວພັນ, ບໍ່ແກ້ວ) ເມື່ອນຳອັກສອນຕົ້ນແບບມາທົດລອງໃນການຝຶກຮູບແບບ Train with existing box ໂດຍການນຳໃຊ້ຊຸດຂໍ້ມູນ Tiff/Box1 ຈະໄດ້ຜົນລັບດັ່ງນີ້ (ເຊກຸອຸງ, ອັດຕຸະປີ, ຫົວຜັນ, ປໍແກ໊ວ), ສ່ວນຮູບແບບ Shape Clustering ທີ່ນຳໃຊ້ຊຸດຂໍ້ ມູນ Tiff/Box2 ແມ່ນໃຫ້ຜົນຮັບແມ່ນ (ເຊກອດິ, ອັດຣີປີ, ຫົວຜັນ, ປົແກ໊ວ). ຈຸດປະສົງຂອງບົດຄົ້ນຄວ້າແມ່ນເພື່ອນຳໃຊ້ເຄື່ອງມືທີ່ບໍ່ມີຄວາມຊັບຊ້ອນໃນການຝຶກຂໍ້ມູນ ແລະ ການກະກຽມຂໍ້ມູນ ແລະ ສະພາບແວດລ້ອມເຫຼົ່ານີ້ແມ່ນຍັງບໍ່ມີງານວິໄຈໃນການຈຳພາສາລາວ ສັງເກດເຫັນວ່າມີຄວາມແມ່ນຍຳສູງເຖິງ 60.32% ແມ່ນເຮັດໄດ້ດີກວ່າຜົນໄດ້ຮັບຂອງຂໍ້ຄວາມແມ່ນຢູ່ໃນລະດັບທີ່ສາມາດອ່ານ, ເຂົ້າໃຈໄດ້ ແລະ ມີຄວາມໃກ້ຄຽງກັບຂໍ້ຄວາມຕົ້ນແບບທີ່ນຳມາຝຶກ.

5. ສະຫຼຸບ

ເຕັກໂນໂລຊີ Optical Character Recognition ຫຼື ການຮູ້ຈຳອັກສອນແມ່ນມີຄວາມທ້າທາຍ ເນື່ອງຈາກ ພາສາລາວມີເອກະລັກ ແລະ ໂຄງສ້າງທີ່ຊັບຊ້ອນ ບໍ່ຄືກັບ ຕົວອັກສອນ ພາສາອັງກິດ ເຊິ່ງການຄົ້ນຄວ້າວິໄຈ ໄດ້ນຳສະເໜີ ການຮູ້ຈຳຕົວອັກສອນ ພາສາລາວ ໂດຍນຳໃຊ້ເຄື່ອງມືການຝຶກດ້ວຍ jTessBoxEditor ແລະ Serack Tesseract Trainer ແລະ ທົດລອງກັບການອ່ານປ້າຍລົດໃນປະເທດລາວທີ່ມີຕົວອັກສອນພາສາລາວ ແລະ ຕົວເລກ ໂດຍການສ້າງໄຟລ Tiff/Box, ການຕັດຄຳ 2 ຮູບແບບ ແລະ ແຕ່ລະຮູບມີການຝຶກແບບ train with existing box ແລະ shape clustering ຈາກຜົນການທົດລອງເຫັນວ່າ ການຕັດຄຳ ໂດຍແຍກແຕ່ລະຕົວພະຍັນຊະນະ-ແຕ່ລະຕົວສະຫຼະ ຫຼື ໂດຍການປະສົມພະຍັນຊະນະ ແລະ ຕົວສະຫຼະ ແມ່ນມີຜົນໄດ້ຮັບທີ່ດີແຕ່ຕ່າງການຝຶກ ສະຫຼຸບໄດ້ວ່າ ສາມາດນຳໃຊ້ Tiff/Box ແບບທີ່ແຍກໂຕພະຍັນຊະນະຈະມີຄວາມໄວ, ງ່າຍໃນການຝຶກ ແລະ ຈັດການຊຸດຂໍ້ມູນໄດ້ດີກວ່າ. ໃນຕໍ່ໜ້າຈະມີການເພີ່ມຊຸດຂໍ້ມູນໃຫ້ຫຼາຍ ແລະ ອ່ານກັບປ້າຍລົດທີ່ຖືກຖ່າຍຈາກຕົວຈິງ ເພື່ອໃຫ້ມີຄວາມແມ່ນຍຳ ແລະ ສາມາດນຳໃຊ້ໄດ້ທົ່ວໄປ.

6. ຂໍ້ຂັດແຍ່ງ

ຂ້າພະເຈົ້າໃນນາມຜູ້ຄົ້ນຄວ້າວິທະຍາສາດ ຂໍປະຕິຍານຕົນວ່າ ຂໍ້ມູນທັງໝົດທີ່ມີໃນບົດຄວາມວິຊາການດັ່ງກ່າວນີ້ ແມ່ນບໍ່ມີຂໍ້ຂັດແຍ່ງທາງຜົນປະໂຫຍດກັບພາກສ່ວນໃດ ແລະ ບໍ່ໄດ້ເອື້ອປະໂຫຍດໃຫ້ກັບພາກສ່ວນໃດພາກສ່ວນໜຶ່ງ, ກໍລະນີມີການລະເມີດໃນຮູບການໃດໜຶ່ງ ຂ້າພະເຈົ້າມີຄວາມຍິນດີທີ່ຈະຮັບຜິດຊອບແຕ່ພຽງຜູ້ດຽວ.

7. ເອກະສານອ້າງອີງ

Phommachanh, S., & Phommachanh, K. (2019). Development of Lao Vehicle License Plate Recognition System Using Tesseract OCR. International Journal of Computer Applications, 178(40), 1-5.

Nam, E. (2018). Understanding the Levenshtein distance equation for beginners. Medium.

Phommachanh, S., & Phommachanh, K. (2020). A Study on Lao Vehicle License Plate Recognition System Using Deep Learning. International Journal of Innovative Technology and Exploring Engineering, 9(3), 2228-2232.

Singh, A., & Singh, S. (2018). Vehicle License Plate Recognition Using Deep Learning and OpenCV. International Journal of Engineering Research & Technology, 7(11), 337-341.

Khan, M., & Khan, M. (2019). A Survey on License Plate Recognition Systems. International Journal of Advanced Computer Science and Applications, 10(2), 498-505.

Ammar, A., Koubaa, A., Boulila, W., Benjdira, B., & Alhabashi, Y. (2023). A multi-stage deep-learning-based vehicle and license plate recognition system with real-time edge inference. Sensors, 23(4), 2120.

Kumar, S., & Singh, S. (2022). Comparative analysis of EasyOCR and TesseractOCR for automatic license plate recognition using deep learning algorithm. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 1-5). IEEE.

Kaur, J., & Singh, S. (2019). License plate recognition using OpenCV and Tesseract OCR on Indian number plates. International Journal of Engineering Research & Technology, 8(12), 1-5.

Li, X., Wang, Z., Wang, Y., & Zhang, H. (2018). Vehicle License Plate Recognition Based on Tesseract OCR Engine. In 2018 37th Chinese Control Conference (CCC) (pp. 8537-8541). IEEE.

Vongpradhip, S., Sudsang, A., & Sinthupinyo, S. (2017). Enhanced Vehicle License Plate Recognition System Using Deep Learning. In 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE. https://doi.org/10.1109/JCSSE.2017.8026139.

Hlaing, Z. T., Zaw, Z., & Oo, K. Z. (2018). Vehicle License Plate Recognition System Based on Tesseract OCR Engine. In 2018 IEEE Conference on Computer Applications (ICCA) (pp. 1-4). IEEE. https://doi.org/10.1109/COMAPP.2018.8377522.

Li, B., & Wang, C. (2019). Research on License Plate Recognition Based on Tesseract OCR. In 2019 5th International Conference on Control, Automation and Robotics (ICCAR) (pp. 320-323). IEEE. https://doi.org/10.1109/ICCAR.2019.8813754.

Yap, W. J., & Tan, C. C. (2019). Vehicle License Plate Recognition System Based on Deep Learning and Tesseract OCR. In 2019 2nd International Conference on Computational Intelligence and Intelligent Systems (pp. 22-26). IEEE. https://doi.org/10.1109/CIIS.2019.8883942.

Li, M., & Xu, Z. (2019). Vehicle License Plate Recognition System Based on Tesseract OCR Engine. In 2019 2nd International Conference on Frontiers of Control and Sensor Networks (CFCSN) (pp. 67-70). IEEE. https://doi.org/10.1109/CFCSN47491.2019.00021.

Khadija El Gajoui, Fadoua Ataa Allah, Mohammed Oumsis. (2015). Training TESSERACT Tool for Amizig OCR. RECENT RESEARCHES in APPLIED COMPUTER SCIENCE: Proceedings of the 15th International Conference on Applied Computer Science (ACS15), Konya, Turkey May 20-22, 2015 (pp.172-179).

Charangan Vasantharajan, Laksika Tharmalingam,Uthayasanker Thayasivam. (2019). Adapting the Tesseract Open-Source OCR Engine for Tamil and Sinhala Legacy Fonts and Creating a Parallel Corpus for Tamil-Sinhala-English.

Chithra Selvaraj,Bhalaji Natarajan. (2018). Enhanced portable text to speech converter for visually impaired.International Journal of Intelligent Systems Technologies and Applications 17(1/2):42.

Kesom, K., & Phawapoothayanchai, P. (2018). Optical Character Recognition (OCR) enhancement using an approximate string matching technique. Engineering and Applied Science Research, 45(4), 282-289.

Beyene, E. G. (2019). Handwritten and Machine printed OCR for Geez Numbers Using Artificial Neural Networks arXiv preprint arXiv:1911.06845. Barkov, A. (2019). The hundred-page machine learning book (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.

ຕາຕະລາງທີ 1. ຜົນການທົດລອງ

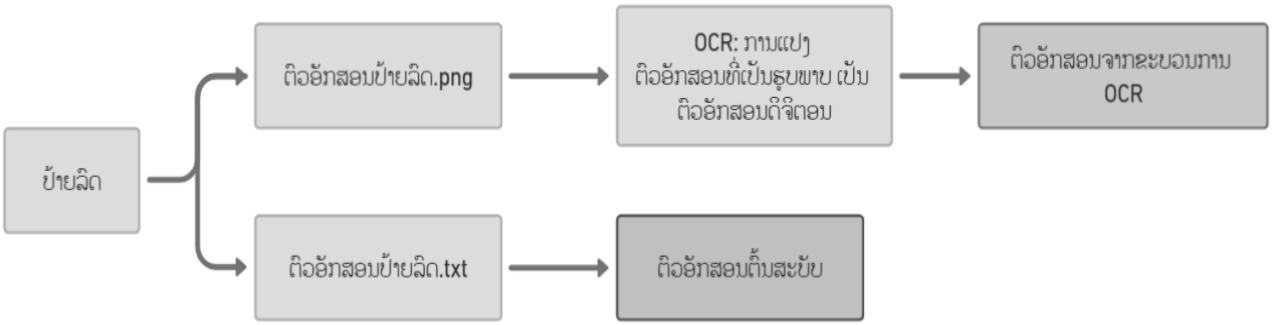
ລຳດັບ [ຮູບປ້າຍລົດ]	ຕົວອັກສອນຕົ້ນສະບັບ	ຕົວອັກສອນ ຈາກຂະບວນການ OCR			
		ການຝຶກແບບທີ1 Tiff/Box1	ການຝຶກແບບທີ1 Tiff/Box2	ການຝຶກແບບທີ2 Tiff/Box1	ການຝຶກແບບທີ2 Tiff/Box2
1	ເຊກອງ ຍປ 6155	ເຊກ່ອງ ຍ່ປ 6155	ດ່ຊກ່ອ່ງ ຍປ 6155	ເຊກອດີ ຍ້ປ 6155	ເຊກອດີ ຍ້ປ 6155
2	ເຊກອງ ຂຜ 5514	ເຊກ່ອງ ຂ່ຜ 5514	ຫ່ຊກອ່ງ ຂຜ 5514	ເຊກຢີ ຂຜ 5514	ເຊກຢີ ຂຜ 5514
3	ຫ້ວຝັນ ຖຝ 8732	ຫ້ວຝ້ນ ຖຝ 8732	ຫ້ອ່ຝ້ນ ຖຝ 8732	ຫ້ວຝ້ນ ຖຝ 8732	ຫ້ວຝ້ນ ຖຝ 8732
. . .	. . .	. . .	. . .	. . .	. . .
19	ສາລະວັນ ຂອ 9345	ສ့ຖອ့ອ້ນ ຂ့ອ့ 9345	ສຖອ့ງ໋ນ ຂອ့ 9345	ສງ໋ວຣ໊ອ້ນ ຂຂ 9345	ສງ໋ວຣ໊ອ້ນ ຂຂ 9345
20	ອັດຕະປື ຍບ 5670	ອ້ດຕ့ຣ໊ປ ຍ့ບ 5670	ອ໋ດຕ့ພ໊ ຍບ 5670	ອ້ດຣ໊ປ ຍ້ບ 5670	ອ້ດຣ໊ປ ຍ້ບ 5670

ຕາຕະລາງທີ 2. ການຝຶກແບບທີ1: train with existing box ໃນTiff/Box 2 ຮູບແບບ

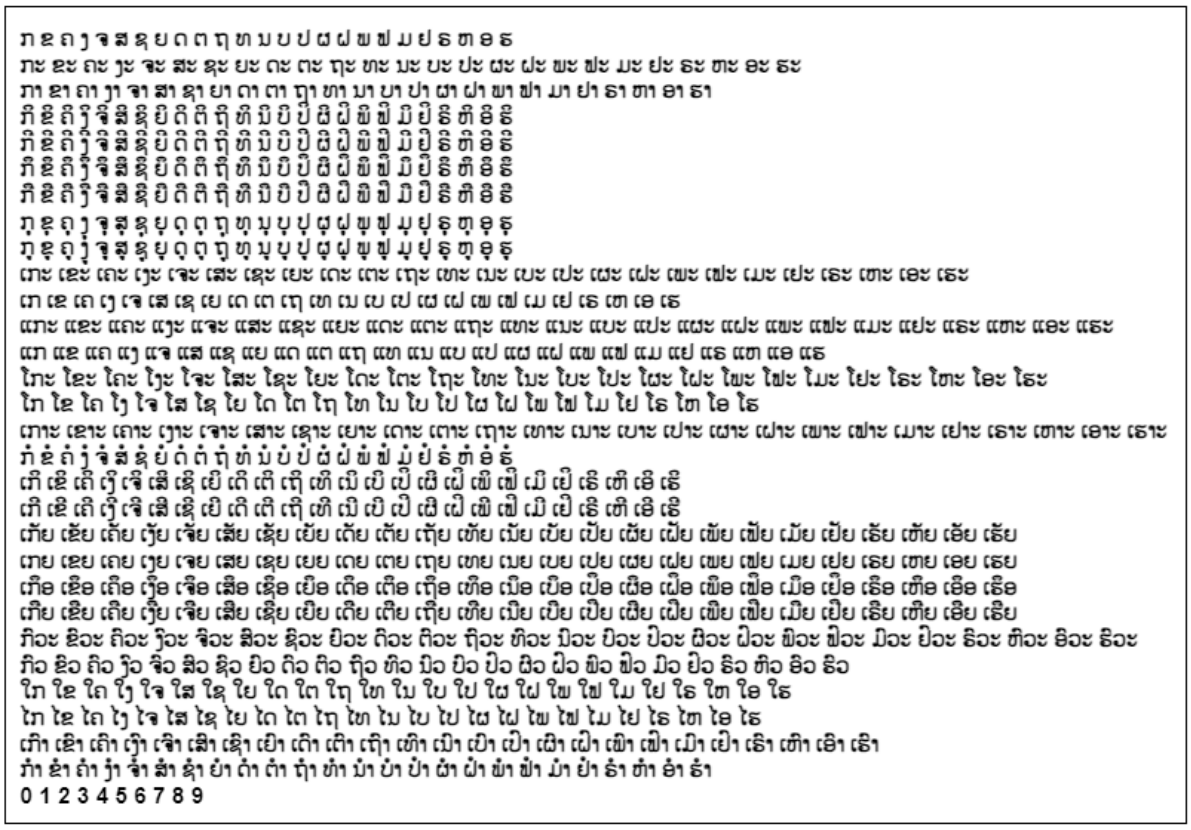
ລຳດັບ [ຮູບປ້າຍລົດ]	Tiff/Box ຮູບແບບທີ1		Tiff/Box ຮູບແບບທີ2	
	Levenshtein Distance	accuracyrate	Levenshtein Distance	accuracyrate
1	6	66.67%	7	61.11%
2	6	66.67%	7	61.11%
3	3	80.00%	6	62.50%
4	6	64.71%	9	52.63%
5	5	72.22%	6	62.50%
6	10	50.00%	13	40.91%
7	9	52.63%	11	42.11%
8	7	61.11%	10	47.37%
9	6	62.50%	8	52.94%
10	8	57.89%	8	52.94%
11	6	66.67%	9	55.00%
12	11	45.00%	13	40.91%
13	8	57.89%	12	42.86%
14	6	64.71%	7	58.82%
15	11	45.00%	12	40.00%
16	5	68.75%	8	52.94%
17	9	43.75%	10	37.50%
18	9	59.09%	12	50.00%
19	9	52.63%	8	52.94%
20	6	66.67%	6	62.50%

ຕາຕະລາງທີ 3. ການຝຶກແບບທີ2: Shape Clustering ໃນTiff/Box 2 ຮູບແບບ

ລຳດັບ [ຮູບປ້າຍ ລຶດ]	Tiff/Box ຮູບແບບທີ1		Tiff/Box ຮູບແບບທີ2	
	Levenshtein Distance	accuracyrate	Levenshtein Distance	accuracyrate
1	5	68.75%	5	68.75%
2	5	66.67%	5	66.67%
3	4	73.33%	4	73.33%
4	5	64.29%	5	64.29%
5	5	68.75%	5	68.75%
6	7	58.82%	7	58.82%
7	9	43.75%	8	55.56%
8	7	58.82%	7	58.82%
9	8	55.56%	7	58.82%
10	8	55.56%	7	58.82%
11	6	68.42%	6	68.42%
12	15	28.57%	15	28.57%
13	6	62.50%	6	62.50%
14	5	66.67%	5	66.67%
15	8	52.94%	8	52.94%
16	5	66.67%	5	66.67%
17	9	43.75%	9	43.75%
18	8	63.64%	8	63.64%
19	9	50.00%	9	50.00%
20	6	64.71%	6	64.71%



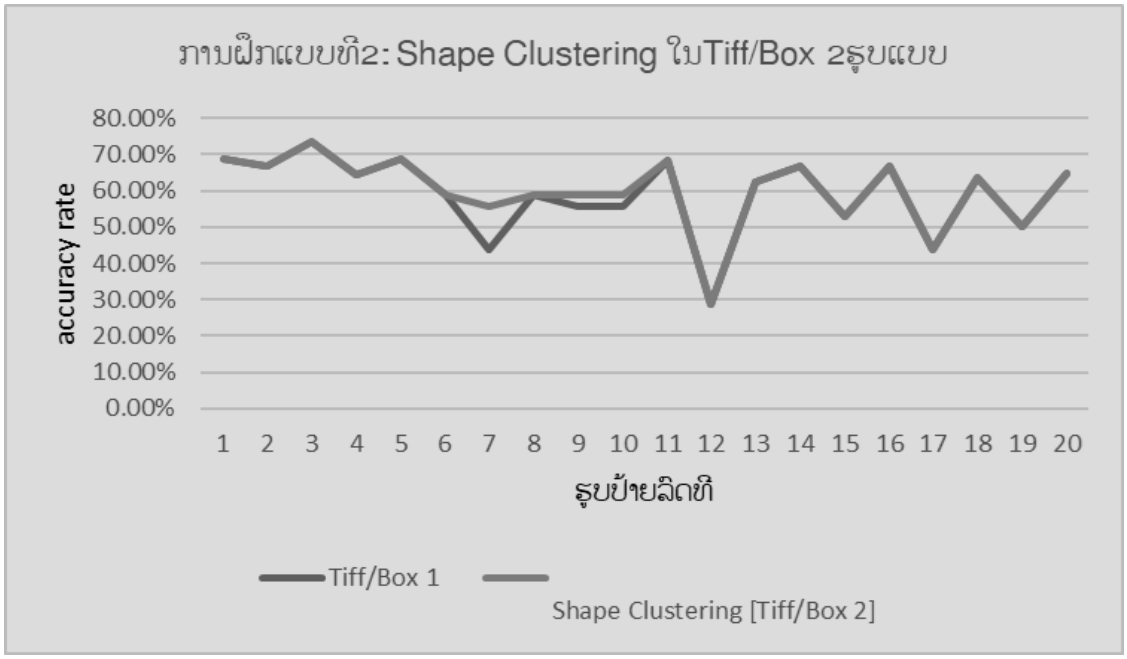
ຮູບພາບທີ 1: ຕົວຢ່າງ ປ້າຍລຶດລາວ



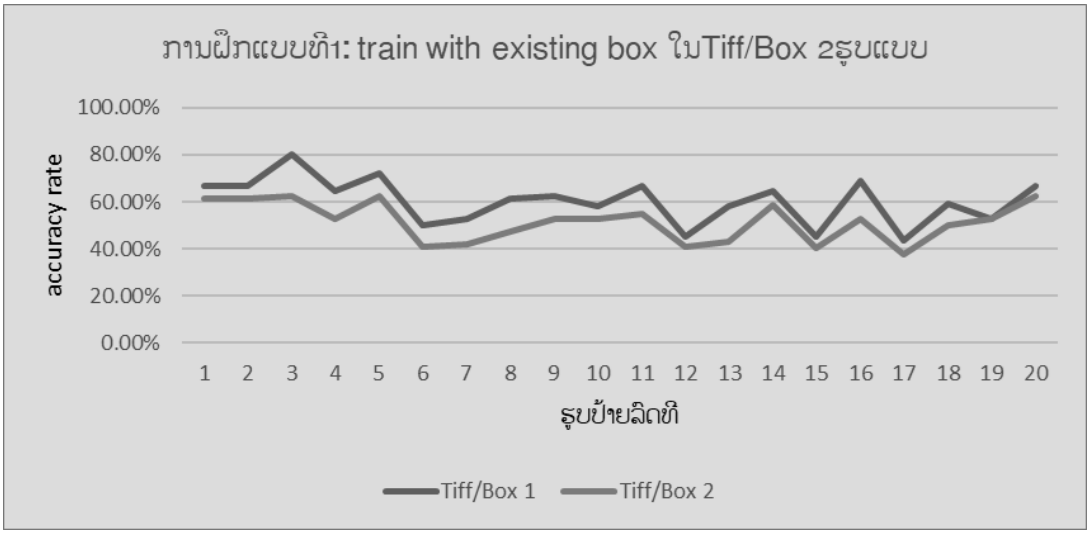
ຮູບພາບທີ 2: ຂັ້ນຕອນການຈັດຈຳຕົວເລກ ແລະ ຕົວອັກສອນ ປ້າຍລົດລາວ

ຮູບພາບທີ 3: ຊຸດຂໍ້ມູນ

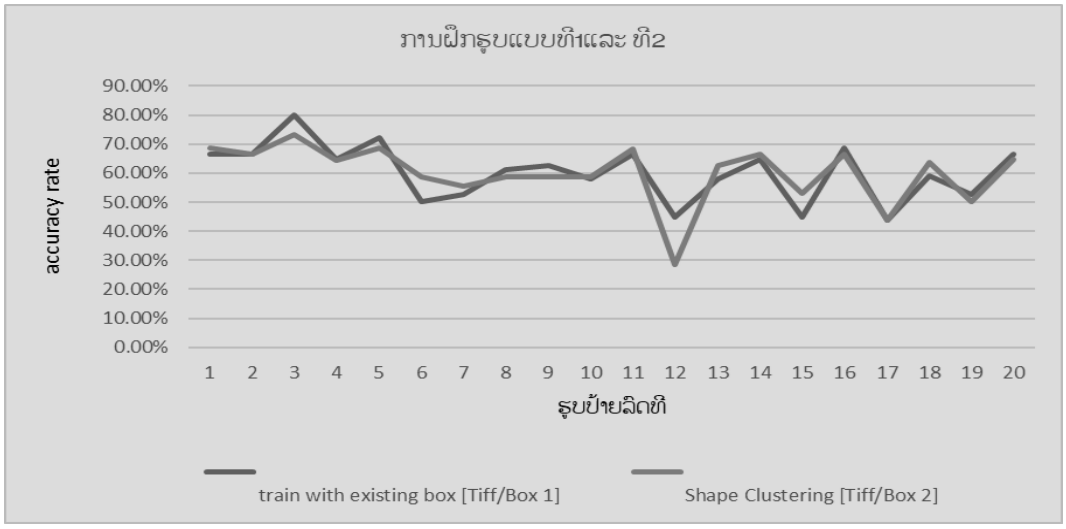
ຮູບພາບທີ 4: ຕົວຢ່າງການສ້າງ Tiff/Box ຕາມຊຸດຂໍ້ມູນ



ຮູບພາບທີ 5: ສະແດງເປີເຊັນຄວາມແມ່ນຍຳການຝຶກແບບທີ1 train with existing box ໃນTiff/Box 2 ຮູບແບບ



ຮູບພາບທີ 6: ສະແດງເປີເຊັນຄວາມແມ່ນຍຳການຝຶກແບບທີ2 Shape Clustering ໃນTiff/Box 2 ຮູບແບບ



ຮູບພາບທີ 7: ສະແດງເປີເຊັນການປຽບທຽມຄວາມແມ່ນຍຳຂອງທັ້ງສອງຮູບແບບ