Title	マルチモーダル深層学習に基づいたタイ手話における指文字認識：新しいベンチマークとモデル構成
Author(s)	WUTTICHAI, VIJITKUNSAWAT
Citation	
Issue Date	2024-06
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19331
Rights	
Description	Supervisor: Nguyen Minh Le, 先端科学技術研究科, 博士

氏　　名	WUTTICHAI VIJITKUNSAWAT		
学位の種類	博士（情報科学）		
学位記番号	博情第 528 号		
学位授与年月日	令和 6 年 6 月 24 日		
論文題目	Deep Multimodal-based Finger Spelling Recognition for Thai Sign Language: A New Benchmark and Model Composition		
論文審査委員	Nguyen Le Minh	JAIST	Professor
	Shinobu Hasegawa	JAIST	Professor
	Naoya Inoue	JAIST	Assoc. Professor
	Kiyoaki Shirai	JAIST	Assoc. Professor
	Masnizah Mohd	Universiti Kebangsaan Malaysia	Assoc. Professor

論文の内容の要旨

Video-based sign language recognition is vital for improving communication for the deaf and hard of hearing. However, due to a lack of resources, creating and maintaining the quality of Thai sign language video datasets is challenging. To address this issue, we assess multiple models with a novel dataset of 90 signs, covering the full letters of alphabets, vowels, intonation marks, and numbers, as demonstrated by 43 signers. We investigate seven deep learning models with three distinct modalities for our analysis: video-only methods (including RGB-sequencing-based CNN-LSTM and VGG-LSTM), human body joint coordinate sequences (processed by LSTM, BiLSTM, GRU, and Transformer models), and skeleton analysis (using TGCN with graph-structured skeleton representation). A thorough assessment of these models is conducted across seven circumstances, encompassing single-hand postures, single-hand motions with one, two, and three strokes, and two-hand postures with static and dynamic point-on-hand interactions. The research highlights that the TGCN model is the optimal lightweight model in all scenarios. In single-hand pose cases, a combination of the Transformer and TGCN models of two modalities delivers outstanding performance, excelling in four particular conditions: single-hand poses, single-hand poses requiring one, two, and three strokes. In contrast, two-hand poses with static or dynamic point-on-hand interactions present substantial challenges, as the data from joint coordinates is inadequate due to hand obstructions stemming from insufficient coordinate sequence data and the lack of a detailed skeletal graph structure. The study recommends integrating RGB-sequencing with visual modality to enhance the accuracy of two-handed sign language gestures. Moreover, experimental results on our dataset show that our method outperforms previous state-of-the-art methods significantly in five out of seven conditional hand pose experiments, especially two-hand poses.

Keywords: Thai Finger Spelling, Sign Language Recognition, Deep Learning, Multimodal Learning, Benchmark Dataset.

論文審査の結果の要旨

The major contribution of the thesis is exploring Thai sign language translation using deep learning. The first significant contribution involves developing a comprehensive video database for Thai Finger Spelling (TFS). This database comprises 10,467 videos demonstrating 90 unique letters in various handshapes with one or both hands, featuring contributions from 43 diverse signers with different appearances and backgrounds.

The second contribution involves extensive research into designing and developing a finger spelling recognizer for TFS based on the collected dataset. The proposed method undergoes analysis through extensive experiments across three modalities and different representation learning techniques. These include CNN-LSTM and VGG-LSTM models for RGB sequences, LSTM, BiLSTM, GRU, and Transformer models for the coordinate sequence of the human body's joint structure, and TGCN for the skeleton modality's graph structure.

Experimental results on various settings and the proposed dataset highlight the advantages and disadvantages of deep learning methods for recognizing Thai Sign Language.

In the third contribution, the thesis conducts comprehensive statistical tests, encompassing both in-sample and out-of-sample evaluations, to identify the most efficient model rigorously. This ensures the recommended model is practical and effective for real-world use. The thesis is presented a nice story for Thai Sign Language which is presented in five well-structured chapters. The publication qualifies for granting the results of the thesis.

Overall, this is an excellent dissertation and we approve awarding a doctoral degree to Wuttichai Vijitkunsawat.