氏　　名	Victor SKOBOV
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2501 号
学位授与の日付	2024 年 3 月 22 日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Sign Language Movement Access and Representation for Linguists and Machines
論 文 審 査 委 員	主　査　坊農 真弓 情報学コース 准教授 PRENDINGER Helmut 情報学コース 教授 相澤 彰子 情報学コース 教授 佐藤 真一 国立情報学研究所 教授 稲邑 哲也 玉川大学 脳科学研究所 教授 中山 英樹 東京大学 大学院情報理工学系研究科 准教授

Summary of Doctoral Thesis

Name in full: Victor SKOBOV

Title: Sign Language Movement Access and Representation for Linguists and Machines

The thesis explores the intersection of linguistics and data science, focusing on the representation and processing of movement in sign language video corpora.

Traditional textual annotation methods in linguistics and computational approaches do not allow for direct access to the captured movement. Linguistic movement notation systems like HamNoSys or SignWriting prove challenging to learn and annotate. Another limiting factor of their use in processing and feature extraction is the lack of adherence to the triangle inequality.

Processing raw high-dimensional frames may lead to the curse of dimensionality. Computer science methods process high-dimensional video data using pre-trained models or pose estimation methods like MediaPipe or OpenPose. Pose estimations provide two-dimensional body key points' estimations on the given frames that are dependent on the camera's position toward the body, which limits the use of the extracted features to the given dataset or camera setting.

The lack of normalized low-dimensional access to the captured movement presents challenges for sign language processing in both fields. The study reviews related work, highlighting the need for a dataset-independent approach. It defines movement representation requirements to address linguistic and data science issues and proposes an appropriate solution. This thesis proposes clearly defined and explainable solutions to overcome these challenges and experimentally demonstrates their benefits.

The first contribution introduces a novel three-dimensional normalization method for MediaPipe's pose estimation, addressing the limitations of two-dimensional projection in capturing a three-dimensional body pose. The method uses the human body proportions to recalculate the third coordinate and rotate the scaled body towards the camera. Lastly, it calculates the directional angles for each joint for movement representation as a series of signals. This normalization method proves effective in enhancing the potential application of resulting models beyond specific camera settings, thus broadening the utility of the collected movement data and allowing feature extraction beyond the captured video corpus. In addition to the technical improvements, the thesis puts forth a human-readable representation of the normalized movement data,

potentially enabling linguists to read the captured movement with the naked eye. That also enables the level of data anonymization and reduces storage requirements.

The methods were proposed to encourage linguistic analysis, demonstrated through an examination of Japanese Sign Language (JSL) sociolinguistic features. The experiments reveal how unique personal movement patterns can be leveraged to identify individual signers, raising concerns about potential threats to anonymity. Notably, the experiments reveal the absence of distinct movement patterns related to signer gender across all data subsets, highlighting consistency across deaf schools, age groups, and prefectures. This finding contrasts with spoken Japanese, where grammatical differences based on gender exist. The classification experiments demonstrate the flaws of projected two-dimensional estimations for sign language processing against the proposed normalization method. The proposed method outperforms basic two-dimensional normalization on a diverse JSL dataset, showcasing its efficacy for machine learning approaches. Results demonstrate the effectiveness of the proposed methods in sociolinguistic feature learning.

To prove the utility of the proposed normalization and representation methods for deep learning approaches, they were tested against the basic normalization in the open sign language recognition benchmark - Word Level American Sign Language. The proposed normalization and encoding demonstrate the ease of preprocessing and dimensionality reduction for dataset samples, making sign language data compatible with speech-oriented conformer models. This combination outperforms other approaches, including multimodal methods, with around a 10% increase in performance compared to basic normalization and three-dimensional coordinates representations.

In conclusion, this thesis offers a comprehensive solution to the technical challenges of sign language processing and movement representation. It emphasizes the collaboration between linguistics and data science, addressing both fields' unique perspectives and requirements. The proposed methods enhance the performance of machine learning models and provide linguists with accessible and processible sign language representations, fostering mutual benefits and paving the way for future advances in both domains.

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　Victor SKOBOV

Title
論文題目　Sign Language Movement Access and Representation for Linguists and Machines

出願者は，2019 年に Google が公開した機械学習を用いたオープンソースである MediaPipe に新しい 3D 正規化法を開発・導入し，手話動作データの精度と適用性を高めてきた．これにより，「単語」単位の手話表現と日本語訳テキストを対象とする従来研究よりもさらに細かい粒度で手話表現の動作の特性分析が可能になった．

本学位論文は，「Sign Language Movement Access and Representation for Linguists and Machines」と題され，全 8 章から構成されている．第 1 章「Introduction to Sign Language Processing」では，手話言語処理の従来研究をデータ収録方法とアノテーションの観点から外観し，言語学的視点・データ科学的視点から問題を整理している．第 2 章「Problem」では，関連する従来手法を整理し，技術的問題を提起している．第 3 章「Methodology」では，本学位論文の技術的な核となる 3D 正規化法を提案している．第 4 章「Experimental Testing: Continious Signs」では，『日本手話話し言葉コーパス』に収録される文単位の連続手話表現に提案手法を適用した結果を示した上で，“Neighbourhood Components Analysis” (NCA)を実施し，日本手話の社会言語学的特徴分析の結果を示している．第 5 章「Experimental Testing: Isolated Signs」では，近年深層学習のベンチマークとして用いられる『単語レベルアメリカ手話データセット』に収録される単語単位の独立手話表現に提案手法を適用した結果，提案手法が従来の基本的な正規化法に比べて約 10%程度精度が向上したことを報告している．第 6 章「Discussion」では，正規化の観点，表現の観点，日本手話の観点，深層学習の観点から議論がなされている．第 7 章「Limitations and Conclusion」では，本提案手法では顔表情を捉えられていないことを課題としてあげ，実験結果が言語学・データ科学の両分野に利点をもたらすことを指摘している．第 8 章「Contribution」では，本学位論文全体をまとめると共に，本研究の技術的・学術的貢献が何かを述べている．

本学位論文には，提案技術を用いた日本手話の社会言語学的特徴の徹底的な分析が含まれている．具体的には，地域，年齢，性別，出身ろう学校などの社会的要因が手話表現の違いに与える影響を分析している．この実証的な分析は，本学位論文が工学研究に閉じず，それ自体に学際的深みと現実世界との関連性を兼ね備え，Google といった企業が開発するオープンソースの実用的な応用と手話言語研究への潜在的な応用可能性を示している．

また同時に，3D 正規化法によって導きだされたデータを人間が読めるように表現する方

法の導入も，本学位論文の独創的な側面である．これは，言語学者が機械学習によって出力されたデータにアクセスできるようにするだけでなく，近年目覚ましい人工知能の技術的な進歩がより広い学術コミュニティで使用可能であることを示すという，広範な研究目標にも貢献している．

公開発表会では博士論文の章立てに従って発表が行われ，その後に行われた論文審査会及び口述試験では，審査員からの質疑に対して適切に回答がなされた．質疑応答後に審査委員会を開催し，審査委員で議論を行った．審査委員会では，出願者の博士研究が新規性が高く学際性豊かであることが高く評価された．以上を要するに本学位論文は，手話言語処理研究における新しい方向性と今後の発展可能性を十分に示したものであり，研究分野の発展に貢献しているという点で学術的価値が大きい．また，本学位論文の成果は，査読付き国際会議 EMNLP(注 1)における Findings(注 2)としてフルペーパー1 件，書籍分担執筆 1 章（一部）として発表され，社会的な評価も得ている．以上の理由により，審査委員会は，本学位論文が学位の授与に値すると判断した．

（注 1）EMNLP は情報学コースが博士論文提出要件の査読付き論文と同等もしくはそれ以上として認める，極めて難関のトップカンファレンスの一つである．
（注 2）Findings はフルペーパー執筆とポスター発表を伴うものである．採択率は低く，競争的な発表カテゴリである．