KIPS Trans. Softw. and Data Eng.
Vol.11, No.8 pp.325~330
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.8.325

Sign Language Dataset Built from S. Korean Government Briefing on COVID-19

Hohyun Sim[†*] · Horyeol Sung[†*] · Seungjae Lee[†††*] · Hyeonjoong Cho[†††]

ABSTRACT

This paper conducts the collection and experiment of datasets for deep learning research on sign language such as sign language recognition, sign language translation, and sign language segmentation for Korean sign language. There exist difficulties for deep learning research of sign language. First, it is difficult to recognize sign languages since they contain multiple modalities including hand movements, hand directions, and facial expressions. Second, it is the absence of training data to conduct deep learning research. Currently, KETI dataset is the only known dataset for Korean sign language for deep learning. Sign language datasets for deep learning research are classified into two categories: Isolated sign language and Continuous sign language. Although several foreign sign language datasets have been collected over time. they are also insufficient for deep learning research of sign language. Therefore, we attempted to collect a large-scale Korean sign language dataset and evaluate it using a baseline model named TSPNet which has the performance of SOTA in the field of sign language translation. The collected dataset consists of a total of 11,402 image and text. Our experimental result with the baseline model using the dataset shows BLEU-4 score 3.63, which would be used as a basic performance of a baseline model for Korean sign language dataset. We hope that our experience of collecting Korean sign language dataset helps facilitate further research directions on Korean sign language.

Keywords : Sign Language Recognition, Sign Language Translation, Sign Language Segmentation, Sign Language Dataset, Deep Learning

대한민국 정부의 코로나 19 브리핑을 기반으로 구축된 수어 데이터셋 연구

심 호 현[†*] · 성 호 렬[†*] · 이 승 재[†††*] · 조 현 중[†††]

요 약

본 논문은 한국 수어에 대하여 수어 인식, 수어 번역, 수어 영상 시분할과 같은 수어에 관한 딥러닝 연구를 위한 데이터셋의 수집 및 실험을 진행하였다. 수어 연구를 위한 어려움은 2가지로 볼 수 있다. 첫째, 손의 움직임과 손의 방향, 표정 등의 종합적인 정보를 가지는 수어의 특성에 따른 인식의 어려움이 있다. 둘째, 딥러닝 연구를 진행하기 위한 학습데이터의 절대적 부재이다. 현재 알려진 문장 단위의 한국 수어 데이터셋은 KETI 데이터셋이 유일하다. 해외의 수어 딥러닝 연구를 위한 데이터셋은 Isolated 수어와 Continuous 수어 두 가지로 분류되어 수집되며 시간이 지날수록 더 많은 양의 수어 데이터가 수집되고 있다. 하지만 이러한 해외의 수어 데이터셋도 방대한 데이터셋을 필요로 하는 딥러닝 연구를 위해서는 부족한 상황이다. 본 연구에서는 한국 수어 딥러닝 연구를 진행하기 위한 대규모의 한국어-수어 데이터셋을 수집을 시도하였으며 베이스라인 모델을 이용하여 수어 번역 모델의 성능 평가 실험을 진행하였다. 본 논문을 위해 수집된 데이터셋은 총 11,402개의 영상과 텍스트로 구성되었다. 이를 이용하여 학습을 진행할 베이스라인 모델로는 수어 번역 분야에서 SOTA의 성능을 가지고 있는 TSPNet 모델을 이용하였다. 본 논문의 실험에서 수집된 데이터셋에 대한 특성을 정량적으로 보이고, 베이스라인 모델의 실험 결과로는 BLEU-4 score 3.63을 보였다. 또한, 향후 연구에서 보다 정확하게 데이터셋을 수집할 수 있도록, 한국어-수어 데이터셋 수집에 있어서 고려할 점을 평가 결과에 대한 고찰로 제시한다.

키워드 : 수어 인식, 수어 번역, 수어 영상 시분할, 수어 데이터셋, 딥러닝

1. 서 론

현재 딥러닝 분야의 수어 연구는 수어 인식, 수어 번역, 수어 영상 시분할의 분야로 나누어져 진행되고 있다. 이러한 딥러닝 수어 연구에 있어서 어려움은 크게 두 가지로 볼 수 있다. 첫째, 손의 움직임, 손의 방향, 표정 등의 정보를 종합적으로 이용해야 하는 수어의 특성에서 오는 어려움이다. 이를

※ 이 논문은 2021년도 정부의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(2021R1F1A1049202).
* 공동 제1저자
† 준 회 원 : 고려대학교 컴퓨터정보학과 석사과정
†† 준 회 원 : 고려대학교 컴퓨터정보학과 석·박사통합과정
††† 종신회원 : 고려대학교 컴퓨터융합소프트웨어학과 교수
Manuscript Received : September 15, 2021
First Revision : November 29, 2021
Accepted : January 1, 2022
* Corresponding Author : Hyeonjoong Cho(raycho@korea.ac.kr)

해결하기 위하여 최근 해외의 수어 인식 연구[1]는 신체의 관절 좌표, RGB 이미지, Optical Flow, Depth 이미지 등 다양한 정보를 이용하여 이를 해결하고자 시도하고 있다. 둘째, 대규모 수어 데이터셋의 부재이다. 현재 해외의 수어 데이터셋[11-16]은 과거부터 현재까지 꾸준하게 수집되고 있으며, 점점 더 많은 양의 데이터셋을 구성하고 과거와는 다른 다양한 수집 방법을 시도하는 등 데이터셋 수집 과정이 구체화 되고 있다. 이에 반해 문장 단위의 한국어-수어 데이터셋은 KETI[2] 데이터셋이 유일한 상황이다. 이에 본 논문은 한국어-수어 데이터셋의 부재를 해결하고자 대규모 수어 데이터셋 수집을 진행하였다. 수어 연구를 위한 데이터셋은 Isolated, Co-articulated 데이터셋 두 가지로 분류할 수 있다. Isolated 수어 데이터셋은 단어, 즉 수어 의미소(글로스, Gloss)를 기반으로 수집된 데이터셋을 일컫는다. 이는 수어 단어에 대한 영상과 텍스트 쌍으로 구성되어 있다. 그러나 수어 연구의 근본적인 목적이 청각 장애인의 의사소통을 개선하기 위해서라면 문장 단위로 수집된 데이터셋이 필수적이다. 의사소통은 단어의 나열이 아닌 수어 문법구조에 따른 문장으로 이루어지고, 같은 문장이라도 수어를 하는 사람마다 조금씩 다른 손의 동작과 표정을 가지기 때문에 문장 데이터의 수집이 필요하다. 이를 위해서 수집되고 있는 데이터셋이 Co-articulated 데이터셋이며, 이는 Continuous 수어 데이터셋이라고도 불린다. Co-articulated 데이터셋은 문장을 기반으로 하여 수집한 데이터셋을 일컫는다. 이는 수어 문장에 대한 영상과 텍스트 쌍으로 구성되어 있으며, 최근에는 수어 문장 내에서 수어 문장을 구성하는 글로스의 시간적 정보에 대한 주석도 수집하고 있다. 현재 공개된 한국어-수어 데이터셋인 KETI 데이터셋은 총 105개의 문장 및 419개의 단어에 대한 수어 영상으로 이루어진 데이터셋으로 소규모 한국어-수어 데이터셋으로 볼 수 있다. 이에 대규모 한국어-수어 데이터셋 수집을 시도하기 위하여, 본 논문은 중앙방역대책 및 중앙재난안전대책의 코로나 브리핑을 이용하여 Co-articulated 데이터셋을 수집하고, 수어 인식을 위한 딥러닝 베이스라인 모델을 활용하여 성능 평가 실험을 진행하였다.

2. 관련 연구

2.1 수어 인식

수어 인식은 수어에서 의미를 전달하기 위해 사용하는 손의 위치, 손의 움직임, 손의 방향, 표정, 몸의 움직임과 같은 시각적 요소를 인식하여 한국어로 변환해주는 것을 말한다. 이러한 수어 인식을 위한 연구로는 글로스 단위의 Isolated 수어 인식과 문장 단위의 Co-articulated 수어 인식으로 나누어진다. 과거부터 활발하게 연구가 되던 Isolated 수어 인식 연구에 비하여 어려운 난이도에 속하는 Co-articulated 수어 인식[3, 4]이 최근 들어 활발히 이루어지고 있다. Isolated-sign 인식은 수어 단어 사전과 같은 단어 기반의 단순한 수어에 이용된다. 실생활 수어는 수어의 문법적 구조에 의하여 문장으로 구성되어 있으며 단순 수어 글로스의 나열로는 볼 수 없다. 실생활 수어와 같은 Co-articulated 수어가 어렵다고 여겨지는 이유는 수어 글로스의 시작과 끝이 전후 수어 글로스에 따라 달라지기 때문이다. 이는 Isolated 수어와 Co-articulated 수어의 결정적인 차이점이라 볼 수 있으며 이러한 점이 Co-articulated 수어 인식에서의 어려운 점이라고 볼 수 있다.

2.2 수어 번역

수어 번역 연구는 sign-gloss-text, sign-text의 두 가지 방향으로 연구가 진행되고 있다. sign-gloss-text 방법은 수어 영상에서 수어 글로스로의 인식 과정, 수어 글로스에서 한국어 문장으로의 번역과정을 파이프라인으로 구성하여 수어 번역을 진행한다. 이러한 방법은 Encoder-Decoder의 구조로 구성되어 있고, Encoder 훈련을 위해 글로스를 Ground-Truth로 활용하는 특징을 갖고 있다[5,6]. 그러나 이러한 방법들은 수어가 가지는 연속적인 특성, 즉 시간적인 정보를 이용하지 못한다는 단점이 있었다. 이를 해결하기 위하여 sign-text 방법으로 수어 영상에서 한국어 문장으로 번역이 진행되고 있다. [7]에서 제시한 방법은 I3D[8] 모델을 이용하여 3D feature를 추출함으로써 시간적 정보를 수어 번역에 이용하였다.

2.3 수어 영상 시분할

수어 영상 시분할 연구는 최근 들어 연구가 조금씩 진행되고 있다. Co-articulated 수어에서 존재하는 수어 글로스가 앞, 뒤 수어 글로스에 따라서 시작 동작과 끝 동작이 달라지는 문제가 수어 영상 시분할 연구의 어려운 점이다. 이를 해결하기 위해서는 수어 문장 영상에서의 수어 글로스 단위의 시간적 정보에 대한 특징 추출이 중요하다. 현재 해외에서의 수어 영상 시분할 연구는 sign spotting, sign segmentation이라는 명칭으로도 진행되고 있으며, attention 기법을 이용한 방법[9], semi-supervised 기법을 이용한 방법[10] 등 다양한 방법이 시도되고 있다.

2.4 수어 데이터셋

수어 데이터셋의 종류는 두 가지로 나누어질 수 있는데, Isolated 수어를 위한 데이터셋과 Co-articulated 또는 Continuous 수어라 불리는 데이터셋이다.

Isolated 수어 데이터셋의 사례는 Table 2에서 볼 수 있다. Isolated 수어는 수어의 동작이 다른 요소에 방해를 받지 않는 수어를 말하며 글로스 단위의 수어를 일컫는다. 이러한 수어는 데이터의 수집 방법이 Co-articulated 수어보다 간편할 뿐만 아니라 주석 또한 쉽게 달 수 있다. 또한, Co-articulated 수어 연구보다 인식 성능에 있어서 비교적 쉬운 분야에 속한다고 볼 수 있다.

Co-articulated 수어 데이터셋의 사례는 Table 3에서 볼 수 있다. Co-articulated 수어는 손의 위치, 손의 방향, 몸의

움직임 등이 앞뒤 수어 글로스에 따라 영향을 받는 수어를 말한다. 즉 문장 단위의 수어 데이터셋을 말하며 Continuous 수어로도 불린다. Co-articulated 수어의 인식은 현재 수어 인식 분야에서도 비교적 도전적인 과제로 여기며 연구가 진행되고 있다. 문장의 구성에 따라 수어 데이터가 달라지기 때문에 Isolated 수어보다 더욱 많은 데이터와 주석을 필요로 한다. Table 2와 Table 3을 비교해 보아도 Isolated 데이터셋이 Co-articulated 데이터셋보다 samples를 더 많이 포함한 것을 볼 수 있다.

3. 대한민국 정부의 코로나 19 브리핑을 기반으로 구축된 수어 데이터셋

본 논문이 수집한 데이터셋은 중앙방역대책본부의 2020년 2월 23일의 브리핑부터 2020년 7월 2일까지의 브리핑, 중앙재난안전대책본부의 2020년 2월 25일의 브리핑부터 2020년 7월 10일까지의 브리핑을 통하여 수집하였다. 대한민국 정부의 코로나 19 브리핑을 기반으로 구축된 수어 데이터셋이라 볼 수 있으며 이를 영문으로 Sign Language Dataset built from Korean Government Briefing on Covid-19, 즉 축약하여 SKBC 데이터셋이라고 부르기로 하였다. 수집한 영상들은 대략 30분 길이를 가지며, 텍스트는 속기자료를 이용하였고 Table 1의 SKBC 부분과 같이 문장별로 번호순으로 정렬하여 사용하였다. 또한, 모든 데이터는 기자 Q&A를 제외한 내용을 수집하였다. 이는 Q&A 구간이 잘 정리되어 발표된 브리핑 구간보다 수어 영상과 문장의 맵핑에 있어 노이즈가 많기 때문이다. 수집한 영상 데이터는 전문 수어 통역사를 통하여 속기자료의 문장 단위로 segmentation 과정을 수행하였다. 이후 영상 속 수어 통역사의 부분을 400x600 크기로 크기를 조정하여 데이터셋을 구성하였다.

본 데이터의 특성은 세 가지로 볼 수 있다. 첫째, 총 11,402개의 문장에 대한 수어 영상 및 문장에 대한 데이터를 수집하여 기존 한국어-수어 데이터셋에서는 존재하지 않던

Fig. 1. Domain of Dataset

대규모의 수어 문장 데이터셋을 구성하였다. 둘째, 문장 영상과 텍스트가 쌍으로 존재하여 Co-articulated 수어 인식 및 번역을 위한 weakly annotation이 존재하는 데이터셋으로 볼 수 있다. 셋째, Fig. 1과 같이 코로나 및 방역에 대한 도메인을 가진 수어 데이터셋이라고 볼 수 있다. 데이터셋의 구성은 Table 4와 같이 구성되어 있다.

4. 베이스라인 모델 실험 및 결과

데이터셋의 성능 평가를 위하여 수어 번역에서 뛰어난 성능을 보인 TSPNet[7] 모델을 이용하였다. TSPNet 모델을 선정한 이유는 다음과 같다. 이 모델은 RWTH-PHOENIX-Weather 2014 T(RPWT) 데이터셋에 대하여 State Of The Art(SOTA)의 성적을 가지고 있다. RPWT 데이터셋은 수어 비디오, 글로스, 문장이 병렬로 구성된 말뭉치 데이터셋이며, 3년간의 독일의 일간 뉴스 및 일기 예보 방송에 관한 일관된 내용으로 구성되어 있다는 점에서 SKBC 데이터셋과 비슷한 성격을 가진다. 이에 RPWT에 대한 SOTA 성능을 가진 TSPNet을 베이스라인 모델로 선정하였다.

Table 1. Comparison of Korean Sign Language Datasets's Sentence Examples

Datasets	ID	Sentence
KETI[2]	1	화상을 입었어요.
	2	폭탄이 터졌어요.
	3	친구가 숨을 쉬지 않아요.
	4	집이 흔들려요.
	5	집에 불이 났어요.
SKBC(ours)	1	코로나바이러스감염증-19 국내 발생현황을 말씀드리겠습니다.
	2	코로나19 증상으로 의료기관 방문 전에는 반드시 콜센터, 보건소를 문의하시고 선별진료소를 우선 방문하여 진료 받으실 것을 권고 드립니다.
	3	어제 신규 확진자는 74명이었고, 격리해제자는 303명 증가하여 전체적으로 격리 중인 환자 수는 감소하고 있는 상황입니다.
	4	어제 서울 지역에서는 6명의 신규 환자가 발생하였습니다.
	5	유럽 입국자는 검역 및 방역당국의 조치에 충실히 따라주시기를 바랍니다.

Table 2. Overview of Isolated Sign Language Datasets

Isolated Datasets	language	Signs	Signers	Samples
ASLLVD[11]	American	2,742	6	9,794
DEVISIGN[12]	Chinese	2,000	8	24,000
MSASL[13]	American	1,000	222	25,513
WLASL[14]	American	2,000	119	21,083
AUTSL[15]	Turkish	226	43	38,336

Table 3. Overview of Co-articulated Sign Language Datasets

Co-articulated Datasets	language	Signs	Signers	Samples
RPWT[16]	German	1,231	9	8257
SIGNUM[17]	German	450	20	15,600
S-pot[18]	Finnish	1,211	5	5,539
BSL Corpus[19]	British	5K	249	40,000
KETI	Korean	419	6	14,672
SKBC (ours)	Korean	4,467	6	11,402

Table 4. Composition of the Dataset

Property	Description
Number of sign sentences	11,402
Number of signers	6
Total samples	11,402
Modality	RGB
RGB resolution	400x600
FPS	30

4.1 베이스라인 모델

수어의 의미는 단일 프레임이 아니라 영상 속 10에서 20 프레임 동안의 손의 방향, 모양, 몸의 동작과 표정 등의 복합적인 정보를 통해 나타난다. 이는 수어에서의 공간적 정보뿐만 아니라 시간적 정보 또한 중요하다는 것을 알 수 있다. 베이스라인 모델로 사용한 TSPNet은 기존 단일 프레임에 대한 주석이 필요한 CNN 모델들[20, 21]의 사용을 넘어, 비디오 즉 여러 프레임에 대한 시간적, 공간적 의미를 모두 사용하도록 시도한 연구이다.

TSPNet은 총 다섯 가지의 과정으로 수어 번역을 시도하였다. 첫째, 입력 데이터를 같은 길이의 비디오 클립으로 나누어 주며 이는 pivot segment라 부른다. 또한, 각 클립은 중복되도록 비디오 segment를 생성한다. 둘째, pivot segment와 인접한 프레임의 segment로 이루어져 있는 surrounding neighborhood를 segment를 구성한다. 이후 surrounding neighborhood segment의 길이를 늘여가며 반복 구성하여 Semantic Pyramid를 만들어낸다. 셋째, 각각의 비디오 segment들에 대하여 3D convolution network 중 하나인 I3D 모델을 이용하여 피쳐를 추출한다. 넷째, inter-scale attention, 즉 계층 내부에서의 attention을 진행하여 지역적인(local) 의미의 일관성을 강화한다. 마지막으로 intra-scale attention, 즉 계층 간의 attention을 진행하여 비지역적인(non-local) 모호성을 해결한다. 이후 디코더를 통하여 수어 영상을 번역한 결과인 텍스트를 추론해내게 된다.

4.2 실험 및 결과

SKBC 데이터셋에 대하여 베이스라인 모델을 사용하여 실험을 진행하였다. 모델 성능 평가 지표로 BLEU Score를 사용하였으며, 3D 피쳐 추출을 위하여 3D 피쳐 추출 모델인 I3D 모델에 대하여 WLASL 데이터셋과 MS-ASL 데이터셋을 이용하여 먼저 fine tuning을 진행하였다. 이후 SKBC의 형태인 수어 영상-수어 문장 형태에 맞게 데이터셋들을 사용하여 RPWT와 SKBC의 비교 실험을 진행하였다. Table 5는 SKBC 데이터셋의 테스트셋과 RPWT 데이터셋의 테스트셋에 대한 성능표이다.

RPWT 데이터셋이 BLEU-4 Score에 대하여 13.21의 성능을 가지는 것에 반하여 본 논문의 데이터셋은 3.63의 성능을 얻었다. 이는 데이터셋의 구성에서 오는 차이와 한국어의 특성에 따른 차이라 볼 수 있다. Table 3에서 볼 수 있듯이 RPWT 데이터셋은 총 8,257문장에 대하여 총 1,231개의 수어를 가진다. 이에 반해 SKBC 데이터셋은 11,402문장에 대하여 총 4,467개의 수어를 가지는 것을 볼 수 있다. 이는 RPWT 데이터셋이 대략 수어 글로스마다 8번의 빈도를 가진다고 볼 수 있고, SKBC 데이터셋은 대략 수어 글로스마다 3번의 빈도를 가지는 것을 알 수 있다. 이는 빈도의 평균이며 실제 데이터셋 내에서의 글로스의 자세한 빈도는 Table 6에서 볼 수 있다. 수어 글로스의 10번 이상의 빈도가 번역 모델 학습을 원활하게 진행하도록 해줄 것에 반해, 총 4,665개의 수어 중 3,359개의 수어 글로스가 부족한 빈도를 가지고 있었다. 또한, 이러한 이유는 SKBC 데이터셋이 코로나 브리핑에 대한 일정 기간의 데이터를 수집한 것에 있다. 이는 기간별 코로나 이슈가 달라짐에 있어서 다양한 상황 및 시간대를 나타내는 문장이 다수 포함되었으며, 이것이 낮은 빈도수를 가지는 단어들이 생기게 했다. 또한, 한국어가 형태소로 이루어져 있다는 점이 다른 언어의 번역보다 어렵게 하였고, 이러한 어려움이 성능 저하의 원인이 되었다. 즉, 향후 SKBC를

Table 5. Experimental Results on Two Datasets

Datasets	BLUE-1	BLUE-2	BLUE-3	BLUE-4
RPWT	35.81	23.02	16.78	13.21
SKBC(ours)	9.18	6.08	4.56	3.63

Table 6. The Frequency of Words of SKBC Datasets

Frequency	Number
300 〉=	85
100 〉=	184
50 〉=	198
10 〉=	839
10 〈	3,359

보강하거나 한국 수어 데이터셋을 새롭게 구축할 때, 문장에서의 글로스별 빈도를 일정값 이상 확보하는 것이 베이스라인 모델의 인식 성능을 높일 수 있음을 의미한다.

5. 결 론

본 논문에서는 딥러닝 연구를 위한 한국어-수어 데이터셋 수집을 위하여 중앙방역대책본부 및 중앙재난안전대책 본부의 코로나 브리핑 자료를 수집하였다. 총 4,665개의 수어 글로스, 11,402문장으로 기존에는 존재하지 않던 대규모의 한국어-수어 데이터셋을 수집하였다. 각 데이터의 구조는 브리핑 자료의 문장별로 segmentation 되어 있기 때문에 Co-articulated, Continuous 수어 인식, 수어 번역 및 수어 영상 시분할을 위한 연구로 사용할 수 있을 것이다. 본 논문의 데이터셋에 대한 성능을 평가하기 위하여 본 논문의 데이터셋과 유사한 성격을 가졌다고 생각하는 RPWT 데이터셋에서 SOTA의 성적을 거둔 TSPNet을 베이스라인 모델로 사용하여 실험을 진행하였다. 실험을 통하여 BLEU-4 Score에서 3.63이라는 값을 얻게 되었고, RPWT 데이터셋의 베이스라인 모델에 대한 성능에 비해서는 낮은 값임을 보였다. 이에 대한 고찰을 통해 값의 차이가 데이터셋이 가지는 수어 글로스 별 문장에서의 빈도수 차이에서 비롯되었음을 지적하였다. 즉, RPWT는 각 수어 글로스가 평균 7, 8개의 문장에서 나타나지만, SKBC는 각 수어 글로스가 평균 3, 4개의 문장에서 나타나고, 데이터셋을 살펴본 결과 수어 글로스가 10번 이상 문장에서 나타나는 글로스가 총 1306개로, 데이터셋 내에 존재한 총 수어 글로스의 약 27%임을 확인하였다. 또한, 코로나의 발생이 매번 다른 공간 다른 시간에서 이루어짐으로 인해 데이터셋 내에서 빈도가 높은 단어와 빈도가 낮은 단어의 구분이 생기게 되었다. 이러한 데이터의 불균형이 모델 학습 및 성능에 영향을 주었으며 RPWT 데이터셋의 성능과 차이를 만들게 되었다. 향후 이러한 문제점은 현재도 매일 축적되어가고 있는 코로나 브리핑 자료에 대한 더 많은 양의 데이터 수집을 통해서나, 또는 각 문장의 수어 글로스 구성을 계산하여 수어 글로스의 반복 빈도가 균일하게 높아지도록 데이터 수집을 한다면 해결할 수 있을 것이다. 본 연구의 경험을 바탕으로 후속 연구를 진행한다면 더 정교하고 방대한 한국어-수어 데이터셋의 구성을 시도할 수 있을 것이다.

References

[1] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Applied Sciences*, Vol.9, No.13, pp.2683, 2019.

[3] S. Russell and P. Norvig, "Artificial intelligence: A modern approach," 3th ed., New York: Prentice Hall, 2009.

[4] J. L. Hennessy and D. A. Patterson, "Instruction-level parallelism and its exploitation," in Computer Architecture: A Quantitative Approach, 4th ed., San Francisco, CA: Morgan Kaufmann Pub., ch.2, pp.66-153, 2007.

[5] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden "Neural sign language translation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*," 2018.

[6] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," In *Advances in Neural Information Processing Systems*, Vol.33, pp.12034-12045, 2020.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on* Computer Vision and Pattern Recognition (CVPR), 2017.

[9] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Read and attend: Temporal localisation in sign language videos," *arXiv preprintarXiv:2103.16481*, 2021.

[10] K. Renz, N. C. Stache, N. Fox, G. Varol, and S. Albanie, "Sign segmentation with changepoint-modulated pseudo-labelling," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[11] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the American sign language lexicon video dataset (ASLLVD) corpus," in *5th Workshop Represent. Processing of Sign Languages: Interactions between Corpus Lexicon (LREC)*, 2012. [Internet], https://open.bu.edu/handle/2144/31899.

[12] X. Chai, H. Wanga, M. Zhoub, G. Wub, H. Lic, and X. Chena, "DEVISIGN: Dataset and evaluation for 3D sign language recognition," Beijing, China, *Technical Report*, 2015.

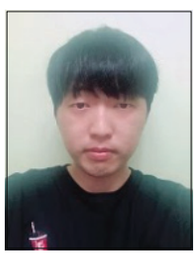
[13] H. R. V. Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding American sign language," 2018, arXiv:1812.01053. [Internet], http://arxiv.org/abs/1812.01053.

[14] D. Li, C. R. Opazo, X. Yu, and H. Li, ''Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,'' in *Prodeedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp.1459-1469, 2020.

[15] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, Vol.8, pp.181340-181355, 2020.
[16] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather," In *International Conference on Language Resources and Evaluation (LREC)*, 2014.
[17] U. von Agris and K.-F. Kraiss, "Towards a video corpus for signer-independent continuous sign language recognition," In *Proceedings of the 7th Intl. Workshop on Gesture in Human-Computer Interaction and Simulation*, May 2007.
[18] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen, "Spot – a benchmark in spotting signs within continuous signing," In: LREC. 2014.
[19] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, and K. Cormier, "Building the british sign language corpus," *Language Documentation & Conservation*, Vol.7, pp.136-154, 2013.
[20] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," In *Proceedings of the International Conference on Learning Representations*, 2014.
[21] C. Szegedy, et al. "Going deeper with convolutions," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.

심 호 현
https://orcid.org/0000-0003-3161-3275
e-mail : tlaghgus0425@korea.ac.kr
2021년 고려대학교 컴퓨터융합소프트웨어학과(학사)
2021년 ~ 현 재 고려대학교 컴퓨터정보학과 석사과정
관심분야 : 수어 인식, 수어 번역, 수어 영상 시분할

성 호 렬
https://orcid.org/0000-0002-8132-4183
e-mail : supernova817@korea.ac.kr
2021년 고려대학교 컴퓨터정보학과(학사)
2021년 ~ 현 재 고려대학교 컴퓨터정보학과 석사과정
관심분야 : 수어 인식, 수어 번역, 수어 영상 시분할



이 승 재
https://orcid.org/0000-0003-3828-3539
e-mail : zaxcv@korea.ac.kr
2018년 고려대학교 컴퓨터정보학과(학사)
2018년 ~ 현 재 고려대학교 컴퓨터정보학과 석·박사통합과정
관심분야 : Text Reasoning, Machine Learning, Natural Language Processing

조 현 중
https://orcid.org/0000-0003-1487-895X
e-mail : raccho@korea.ac.kr
1996년 경북대학교 전자공학부(학사)
1998년 포항공과대학교 전자전기공학(석사)
2006년 미국 버지니아 공과대학교 컴퓨터공학(박사)
2009년 ~ 현 재 고려대학교 컴퓨터융합소프트웨어학과 교수
관심분야 : Machine Learning Techniques, Action/Gesture Recognition, Sign Language Translation, Application Using Optimization Theory, Cyber-physical Systems