박사학위 청구논문

지도교수 윤경로

심층 신경망을 이용한
한국 수어 단어 인식

2020년 8월

건국대학교 대학원

컴퓨터 · 정보통신공학과

배효철

심층 신경망을 이용한
한국 수어 단어 인식

Word Level Korean Sign Language Recognition
using Deep Neural Network

이 논문을 공학 박사학위 청구논문으로 제출합니다.

2020년 4월

건국대학교 대학원
컴퓨터 · 정보통신공학과
배효철

배효철의 공학 박사학위 청구논문을 인준함.

심사위원장 (인)

심사위원 (인)

심사위원 (인)

심사위원 (인)

심사위원 (인)

2020년 6월

건국대학교 대학원

목차

표 목차

그림 목차

ABSTRACT

Word Level Korean Sign Language Recognition using Deep Neural Network

Hyo-Chul Bae

Department of Computer, Information & Communication Engineering

Graduate School of Konkuk University

Sign language is one of the languages used by people with hearing impairments. However, it can be quite challenging for the hearing impaired to communicate with the public using sign language. This challenge is being addressed in multiple countries with the aid of cutting-edge research. However, sign languages have not been sufficiently researched in Korea. As a result, Korea does not have a dataset for sign language research. A sign language dataset is essential for sign language research. In this study, a novel dataset and a deep learning network model for Korean sign language recognition is presented. The proposed KU Korean Sign Language dataset comprises 1,151 videos recorded by 10 singers with 41 words. These videos emphasize the themes of greeting and conversations frequently used in our daily life. Three data augmentation methods are used to reinforce the insufficient data; the results obtained with these methods were

excellent. In addition, a word recognition model network for Korean sign language word recognition is proposed.

The accuracy of sign language recognition is increased by using a skeleton image. An additional 2% increase is observed by employing 2D convolution for the word recognition network. Thus, the proposed KU Korean sign language dataset and word recognition network exhibit a high word recognition rate of 87%.

The remainder of this paper is organized as follows. Chapter 2 details the existing sign language datasets of multiple countries. In addition, it discusses the existing techniques used in motion recognition and sign language research. Chapter 3 describes the proposed dataset for Korean sign language, as well as a novel deep learning network for its recognition. Chapter 4 examines the results obtained by conducting measurements and experiments on the size of the dataset and input image by using various techniques. Finally, Chapter 5 discusses the scope for future improvements with the help of a study on Korean sign language recognition.

Keyword(9P) : 3D CNN, Korean Sign Language Word Data, Deep Learning, Korean Sign Language

제1장 서론

제1절 연구의 배경

전국의 청각 및 언어 장애를 가지고 있는 사람은 2018년의 조사에 따르면 363,326명에 달하며, 이는 총 장애인의 14%에 해당한다[2]. 또한 청각장애인은 선천적인 장애보다 질환 및 사고와 같은 후천적인 장애로 인한 원인이 더 많기 때문에 청각장애인의 수는 추후에도 지속적으로 증가할 것으로 보인다. 이러한 청각장애인들이 의사소통을 위해 가장 높은 비율로 사용하는 것은 수어이며, 가장 낮은 비율로 사용하는 것은 필담이다. 하지만 청각장애인들이 직장, 관공서 및 금융기관 등의 일상생활을 하면서 가장 많이 사용하는 의사소통 방법은 필담으로 조사됐다.[1] 이는 청각장애인들이 가장 많이 사용하는 의사소통 방법인 수어를 일상생활에서는 거의 사용을 하지 못한다는 의미가 되며, 이로 인해 대다수의 청각장애인들은 의사소통에 큰 어려움을 겪고 있다[1]. 이는 청각장애가 없는 사람들이 수어에 대해 잘 모르기 때문이다. 따라서 수어를 자동으로 분석하고 번역해주는 연구가 필요하다.

세계의 여러 나라에서는 이미 수어를 위한 다양한 연구가 진행되어 왔다. 수어는 지화와 수화로 나뉘게 된다. 지화는 손의 모양으로 알파벳, 자음, 모음, 숫자 등을 표현하는 방법이고, 수화는 손의 동작 및 얼굴 표정을 이용하여 명사 및 동사 등을 표현하는 방법이다. 지화를 쓰는 속도나 읽는 속도가 수화에 비교할 때 상대적으로 느리기 때문에 지화를 이용하여 하는 의사소통에는 어려움을 느끼게 된다. 따라서 통상적으로 지화는 수화단어가 없거나 특정 수화 단어를 상대방이 모르는 한정된 경우에 수화를 대체하여 사용된다. 이러한 수어를 컴퓨터가 이해하기 위해서는 연속된 손의 동작 및 얼굴표정을 이해하여야 하기 때문에 높은 수준의 시공간적 이해가 필요하다. 이를 위해 RGB 카메라를 이용하여 손의 모양을 추출하는 방법[4], 깊이

카메라를 이용하는 방법[16], 장갑을 끼고 손을 인식하는 방법 등 다양한 방법이 시도되어 왔다.[5]. 하지만 여전히 현재 수준의 컴퓨터 비전 및 기계 학습에서는 매우 어려운 문제로 여겨지고 있다[9, 15, 16, 17, 18, 19, 20]. 이러한 이유로 수어에 대한 연구는 실생활에 적용되지 못하고 연구에만 그쳐왔다. 또한 손으로 하나의 문자를 나타내는 지화와[21]는 달리 수화는 손 모양, 손의 동작, 표정에 의해 각 의미가 결정이 된다[22].

최근에는 다양한 딥러닝 기법이 연구되고 또한 검증되면서 세계 여러 나라에서 수화에 대한 연구들이 다시 진행되고 있으나, 이러한 신경망을 이용한 수어 연구를 하기 위해서는 적절한 데이터 세트가 부족한 상황이다. 대부분의 수어는 손의 모양, 손의 위치 및 표정의 연속적인 포즈로 구성된 일련의 동작으로 각각의 의미를 구분한다. 이는 사람이 표현하는 수 많은 단어만큼 수어도 그에 해당하는 많은 동작들이 존재한다는 것을 의미한다. 하지만 현재 존재하는 수어 데이터 세트들은 이런 수많은 단어들을 충분히 수집하지 못하고 있는 실정이다. 수어의 연구를 위해 많은 사람들이 여러 방법들을 통해 다양한 데이터 세트를 만들고 또한 연구하고 있다. 미국에서는 Purdue ASL[6]을 시작으로 3300개 이상의 단어를 포함하는 Boston ASLLVD[7] 및 애니메이션 형태로 만든 CUNY ASL[8]을 만들었고, 가장 최근에는 수어 사용자 100명 이상, 단어 2000개 이상을 포함하는 단어 기반의 WLASL[37]을 발표했다. 미국 수어는 전 세계 20개국 이상의 청각장애인들이 사용하고 있다[39]. 독일에서는 기상수화와 기본적인 단어와 문장을 포함하는 RWTH-PHOENIX-Weather[9] 데이터 세트와 SIGNUM[10]을 공개했다. 중국에서는 각각 RGB, Depth, Skeleton정보를 담고 있는 DEVISIGN-G/-D/-L[11]을, 폴란드에서는 키넥트로 촬영한 PSL Kinect 30[12]과 Time-of-flight카메라로 촬영한 PSL ToF 84[13]를 공개했다. 한국 수어는 전자부품연구원에서 발표한 KETI 수어 데이터 세트와 KSL 데이터 세트가 있다. KETI수어 데이터 세트는 10명의 청각장애인의 105문장, 419단어로 이루어진 총 14,672개의 비디오로 구성되어

있으며[14], KSL 데이터 세트의 경우 20명의 청각장애인이 촬영한 실생활에서 자주 사용되는 77개 단어로 구성되어 있다[3].

제2절 연구의 목표 및 내용

수어는 청각 장애인들이 일상어로 사용하는 하나의 언어이다. 하지만 청각 장애인들이 수어를 이용해 일반인들과 의사소통을 하기에는 매우 어렵다. 전문 수화 통역사의 도움을 받으면 의사소통이 가능하나, 개인의 민감한 사안까지도 모두 통역사를 통해 의사소통을 해야 하는 불편함이 생긴다.

현재 한국 수어는 데이터 세트라고 정의할 수 있는 자료가 많지 않다. 많지 않은 데이터 세트 중 KETI 수어 데이터 세트[14]는 안전과 관련된 데이터 세트로 일상 생활에서 사용하기에는 부족한 부분이 존재하며, 그 데이터의 양 또한 딥러닝 네트워크에서 사용하기에는 부족하다. 또다른 데이터 세트인 KSL 데이터 세트[3]의 경우 영상에 레이블링이 되어 있지 않아 트레이닝 및 검증에 사용할 수가 없다.

본 논문에서는 한국 수어의 인식을 위한 딥러닝 네트워크 모델과 이를 위한 새로운 데이터 세트를 제안한다.

제안하는 딥러닝 네트워크는 한국 수어를 높은 정확도로 인식하기 위한 단어 인식 딥러닝 네트워크 모델이다. 과거의 수어 연구는 지화에 대한 연구위주로 진행이 되었으며, 이는 한국 수어 관련 연구도 비슷하다. 최근 딥러닝 기법이 연구되고 발전되면서 지화가 아닌 수화와 관련된 연구들이 늘어나고 있는 추세이나, 수어의 문장을 인식하기위한 수화와 관련된 많은 연구들은 비교적 낮은 정확도의 인식률을 보여준다[3, 14, 15]. 수어의 문장 인식이 낮은 정확도를 보여주는 이유 중 하나는 낮은 수어의 단어 인식률이

다. 이는 부족한 데이터 세트의 문제도 있지만 더 큰 문제는 네트워크의 구성에 있다. 수어의 인식은 크게 보면 사람의 동작 인식과 매우 유사하다. 대부분의 사람의 동작 인식 관련 딥러닝 연구들은 CNN + LSTM의 구조로 네트워크를 구성 한다[24, 25]. 때문에 대다수의 수어 관련 논문 또한 CNN + LSTM의 구조로 네트워크를 구성한다[3, 14, 20, 23]. 몇몇 연구에서는 성능의 향상을 위하여 기본적인 영상정보 외에 옵티컬 플로우를 함께 사용하거나[26], 사람의 몸의 스켈레톤 영상을 참조하는 등[14, 27]의 다양한 연구들이 진행되어 왔다. 하지만 대부분의 네트워크들은 해당 도메인에 특화되어 있고, 특히 데이터가 부족한 한국 수화에 적용하기에는 상당한 어려움이 따른다.

본 논문에서는 기존의 다양한 동작인식 및 수어 연구에서 사용된 기법들을 조사하고 분석한다. 이를 토대로 수어의 문장을 인식하기 위한 기본 조건인 단어의 인식을 위해 한국 수어에 적합한 단어 인식 네트워크를 제안한다.

또한 이를 위한 데이터 세트를 구축, 제안하는데, 제안하는 데이터 세트는 한국 수어 중 일상 생활 중 사회 생활에서 사용되는 단어 41개를 선정하여 구축하였으며, 부족한 데이터를 위해 데이터 확장 기법을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 수어 연구를 위한 배경지식과 관련된 연구인 여러 국가들의 다양한 데이터 세트 및 기존의 동작인식 및 수어 연구에 사용된 기술들에 대해 살펴본다. 3장은 본 논문에서 제안하는 한국 수어를 위한 데이터 세트에 대한 설명과 새로운 한국 수어 인식을 위한 딥러닝 네트워크를 제안한다. 4장에서는 입력 영상의 크기, 입력 영상의

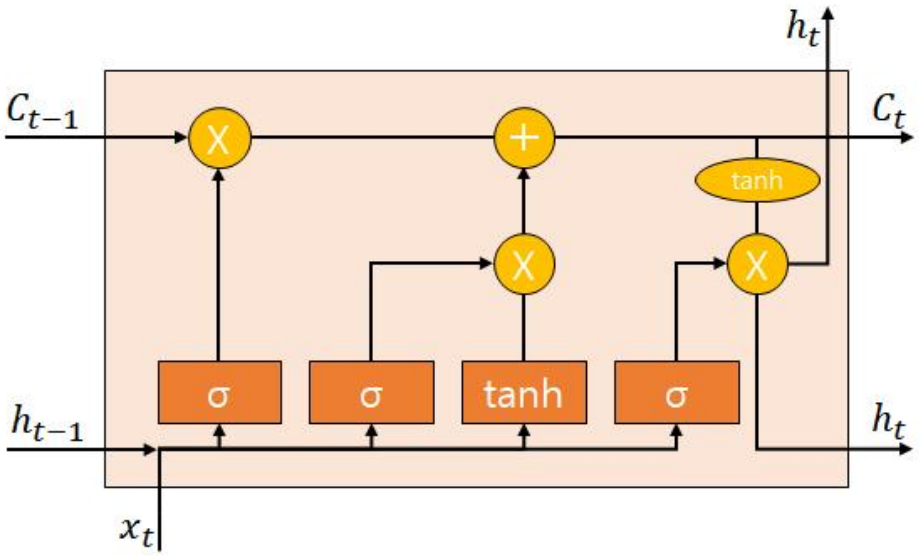
길이, 데이터 증강 및 네트워크 구성에 대한 다양한 방식을 실험해보고 그에 대한 측정 결과에 대해 기술한다. 마지막으로 5장에서는 앞으로 진행되어야 할 한국 수어 인식 연구에 대해 제안한다.

제2장 관련 연구

제1절 시계열 데이터를 위한 딥러닝 네트워크

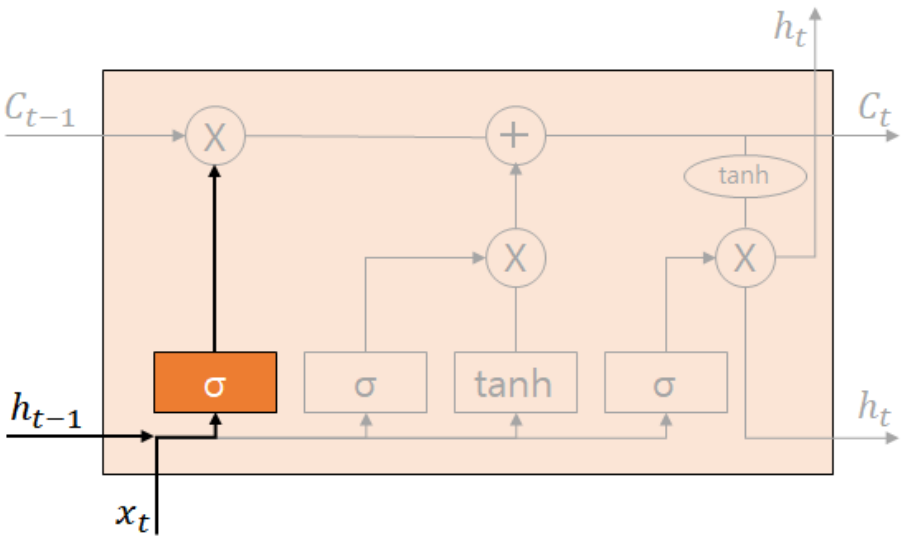
CNN(Convolutional Neural Networks)은 2차원의 영상을 다루고 있다. 이는 이미지넷에서 주최하는 ILSVRC(Large Scale Visual Recognition Competition)대회에서 그 우수성을 보여준다[28, 29, 30, 31]. 하지만 기존의 CNN은 시계열 데이터를 고려하지 않고 있기 때문에, 시계열이 포함된 3차원 영상의 인식을 위해서는 기존의 CNN과는 다른 다양한 연구들이 진행되어 왔다. 본 절에서는 그중 대표적인 LSTM[32]과 3D CNN[33]에 대해 기술한다.

1. LSTM



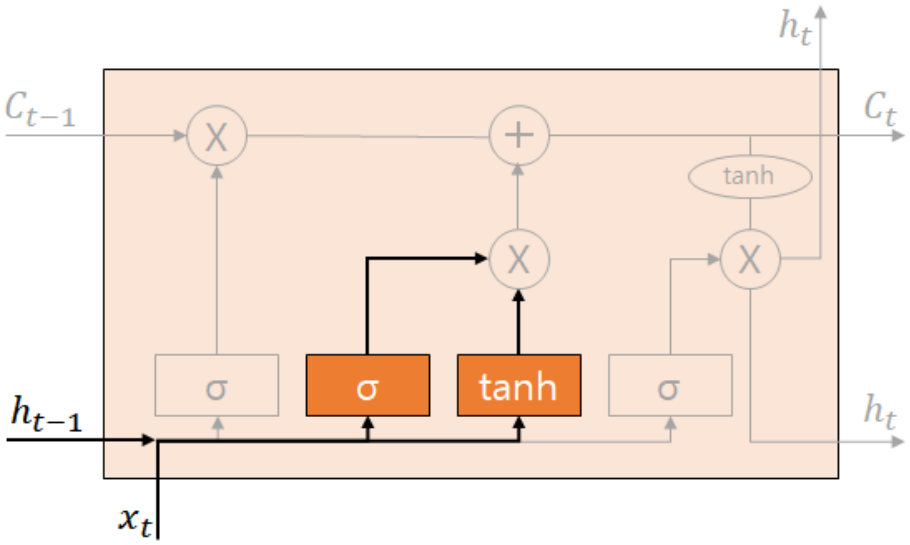
<그림 2-1> LSTM 모듈 구조

LSTM은 RNN(Recurrent Neural Network)[34]의 변형된 모델 중 하나이다. 기존의 RNN은 짧은 시퀀스에 대해서는 괜찮은 효과를 보여주지만, 시퀀스의 길이가 길어지면 길어질수록 앞의 정보가 뒤로 전달이 되지 못하는 현상이 발생한다. LSTM은 이를 극복하기 위해 <그림 2-1>과 같이 명시적으로 설계되었다. <그림 2-1>에서 주황색 박스는 학습된 신경망 레이어를 의미하고 노란색 동그라미는 Pointwise 연산을 의미한다. LSTM은 은닉층의 하나의 모듈에 입력, 망각 및 출력 게이트를 추가하고, 이전 시점과 다음 시점의 상태를 구하기 위한 셀 상태(cell state)를 추가했다. LSTM의 핵심은 바로 이 셀 상태이다. <그림 2-1>에서 가장 위에 그려진 선에 해당하며, 입력, 망각 및 출력 게이트를 통해 셀 상태를 보호하거나 없애는 등의 제어를 한다. 각각의 게이트는 Sigmoid(σ) 레이어와 Pointwise 곱셈으로 구성 되어있다. Sigmoid(σ) 레이어는 0과 1사이의 숫자를 보내며, 그 값에 따라 전달되는 정보를 결정하게 된다.



<그림 2-2> LSTM 망각 게이트 레이어

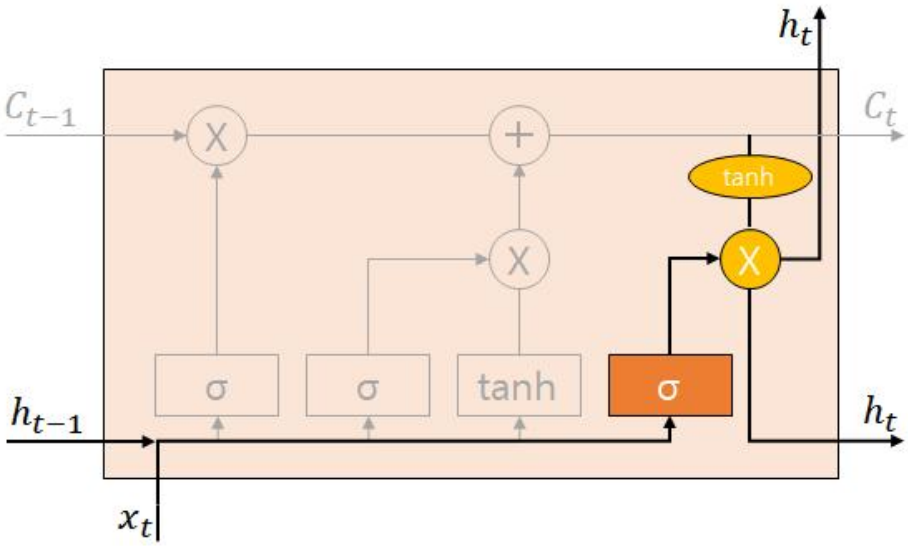
<그림 2-2>는 LSTM 망각 게이트 레이어를 보여준다. LSTM의 첫번째 단계는 셀상태로부터 어떠한 정보를 버릴지를 정하는 것으로, 버리는 정보는 Sigmoid(σ) 레이어에 의해 결정되며 이 단계를 망각 게이트 레이어라 한다. 두번째 단계는 <그림 2-3>에 하이라이트 되어 있는 입력 게이트 레이어다. 먼저 입력되는 새로운 정보들 중 어떠한 정보를 셀 상태에 저장할 것인지를 Sigmoid(σ) 레이어를 통해 결정한다. 그후 tanh 레이어가 새로운 후보 값들인 벡터를 만들어 과거 셀상태인 C_{t-1}을 업데이트 해서 새로운 셀 상태인 C_t를 만들게 된다.



<그림 2-3> LSTM 입력 게이트 레이어

마지막으로 업데이트된 셀 상태에서 어떤 부분을 출력하게 될지를 결정하게 되는데 이를 출력 게이트 레이어라 하며 <그림 2-4>에 하이라이트 되어 있다. 출력 게이트 레이어는 우선 입력 데이터를 Sigmoid(σ) 레이어

가 처리한다. 이때 Sigmoid레이어는 입력된 데이터에 따라 0 또는 1의 값을 내보낸다. 그후 tanh레이어를 통과해 −1과 1사이의 값을 가지는 이전 셀 상태 값과 pointwise 곱셈을 하여 다음 모듈에 셀 상태의 어떤 부분을 전달할지를 결정하게 된다. 이러한 과정을 통해 시퀀스의 길이가 길더라도 LSTM은 RNN보다 앞의 정보를 뒤로 전달할 수 있게 되고 보다 탁월한 성능을 보여준다.

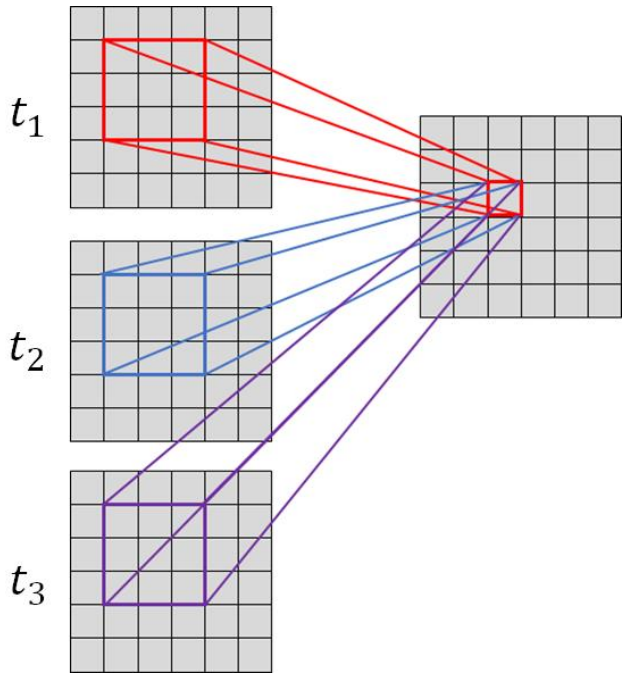


<그림 2-4> LSTM 출력 게이트 레이어

2. 3D CNN

기존의 2D CNN에서 비디오를 분석할 때, 비디오의 각 프레임을 스틸 이미지로 취급하고 이러한 각 프레임에 CNN을 적용하여 동작을 인식한다 [35]. 하지만 이러한 접근법은 다수의 연속된 프레임으로 인코딩 된 모션정보를 고려하지 않고 있다. 이러한 기존의 문제를 해결하기 위해 3D CNN은

공간 및 시간 차원 모두에 따라 구별되는 특징을 포착하도록 구성되어 있다 [33].



<그림 2-5> 3D CNN에서의 합성곱

<그림 2-5>는 3D CNN의 합성곱을 생성하는 과정을 나타낸다. 연속된 3개의 프레임인 t_1, t_2, t_3의 같은 위치로부터 정보를 얻어와 합성곱을 연산한다. 기존의 CNN에서 특징맵을 만들 때, 각 특징맵을 만들 때 사용되는 가중치가 공유되지 않듯이, 3D CNN 또한 가중치가 공유되지 않는 형태로 특징맵이 생성된다.

3D CNN은 위와 같은 구조 덕분에 연속된 프레임으로 구성된 모션정보의 특징들을 추출함에 있어 보다 높은 성능을 보여준다. 하지만 3D CNN은 CNN보다 복잡한 연산을 수행해야 되기 때문에 CNN보다 훨씬 더 많은 자원을 소모하게 되는 단점이 존재한다.

많은 동작인식 및 수어 연구에서 CNN+LSTM의 구조를 선택하거나, 3D CNN을 선택하여 연구를 진행하여 왔다. 많은 결과에서 3D CNN의 성능이 CNN+LSTM구조의 성능보다 높기 때문에[36, 38] 본 논문에서는 3D CNN을 이용한 네트워크 모델을 제안한다.

제2절 동작인식 딥러닝 네트워크

동작인식과 수어의 공통된 점은 연속된 프레임으로 이루어진 비디오라는 점이다. 기존의 CNN으로 처리하던 이미지의 경우 독립된 정지된 영상을 분석하고 처리한 반면, 동작인식의 경우 시간 정보가 추가된 시계열 데이터로 볼 수 있다. 이러한 시계열 데이터를 처리하기 위한 다양한 네트워크들은 네트워크의 구조와 사용하는 데이터에 따라 구별할 수 있다.

1. 구조적 분류

구조적으로 크게 다음과 같이 분류할 수 있다.

- CNN
- CNN + LSTM
- 3D CNN
- 3D CNN + LSTM

CNN은 이미지의 공간적 특징을 추출하는 데는 탁월하나, 시간적 특징은 추출하지 못한다. A. Karpathy는[46] CNN을 이용한 네트워크 모델에서 시계열 데이터를 처리하기 위하여, 단일 프레임만을 학습하는 Single Frame, 영상의 양쪽 끝 프레임을 같이 학습하는 Late Fusion 프레임, 인접한 여러

프레임을 학습하는 Early Fusion 및 인접한 여러 프레임을 여러 위치에서 동시에 학습하는 Slow Fusion등 총 4개의 모델링을 제안하여 네트워크에서 시간적 특징을 포함 할 수 있도록 했다. 그 결과 UCF-101[42] 데이터 세트에서 65.4%의 정확도를 보여줬다.

CNN + LSTM은 영상의 각 프레임에 대해 CNN을 통해 공간적 특징을 추출한 뒤, 추출된 각 프레임의 공간적 특징 값을 기준으로 LSTM을 이용해 시간적 특징을 추출하고, 그 결과를 도출하는 네트워크 모델이다[47]. J. Donahue[47]가 제안한 CNN+LSTM 모델은 동작 인식, 이미지 주석, 동영상 주석 등, 세가지 분야에서 실험을 진행했다. 그 중 동작 인식의 경우 UCF-101 데이터 세트를 사용해 실험했으며, 그 결과는 68.19%의 정확도를 보인다. 이는 CNN만을 사용한 [46]의 결과보다 좋은 결과이다.

3D CNN은 기존의 CNN과는 달리 공간적 특징과 시간적 특징 모두를 추출한다[33]. 이는 CNN+LSTM과 유사하다고 볼 수 있으나, 공간적 특징을 추출하고, 그 뒤 시간적 특징을 추출하는 CNN+LSTM과는 달리 동시에 공간 및 시간적 특징 모두를 추출한다. UCF-101 데이터 세트의 실험 결과는 85.2%의 정확도로 상기 두 방법보다 시공간 데이터를 분석하는데 더 적합함을 보여주었다.

3D CNN + LSTM은 상대적으로 높은 정확도를 보여줄 수 있다[48]. X. Ouyang[48]의 모델을 보면, UCF-101의 데이터셋에 대하여 88.9%의 높은 정확도를 보인다. 하지만 3D CNN 만으로도 엄청난 하드웨어 자원을 필요로 하는데, 추가로 LSTM까지 사용하게 되면 매우 높은 하드웨어 자원이 필요하여 일반적인 고성능 GPU를 활용한 데스크탑 환경에서 사용이 불가능하다는 단점이 있다.

2. 데이터에 따른 분류

두번째로 동일한 네트워크여도 사용하는 데이터의 종류에 따라 정확도가 변경된다. 딥러닝 네트워크에서는 기본적으로 RGB영상을 사용한다. 하지만 RGB영상 뿐만 아니라 옵티컬 플로우, 스켈레톤 영상 등을 사용해서 정확도를 높인 네트워크도 있다.

우선 옵티컬 플로우 데이터를 사용하여 성능을 향상시킨 네트워크로는 K. Simonyan[26], Tran[36], Ng[50] 등이 있다. 3D CNN에서 옵티컬 플로우를 입력 데이터로 사용했을 경우, UCF-101의 정확도는 90.4%로 RGB 영상을 사용했을 때 보다 5.2% 높게 나온다[36]. K.Simonyan[26]의 네트워크의 경우, CNN 네트워크를 사용하며, 입력 데이터로는 RGB로 된 프레임 한 장과 해당 프레임과 주변프레임의 옵티컬 플로우를 사용했다. 그 결과 88%의 정확도를 보여줬다. 이는 기존 CNN을 사용한 A. Karpathy[46]의 슬로우 퓨전보다 무려 22.6% 상승한 결과를 보여준다. J. Yue-Hei Ng[50]의 네트워크는 CNN+LSTM을 사용하였으며, UCF-101 데이터의 실험에서 88.6%의 정확도를 보여줬다.

사람의 스켈레톤 영상을 입력 영상으로 사용한 네트워크는 B. Mahasseni[49]의 LSTM을 이용한 네트워크가 존재한다. 해당 네트워크는 UCF-101 데이터 세트를 사용한 실험에서 86.9%의 정확도를 보여준다. 만약 사람의 스켈레톤 영상에 LSTM 보다 시계열 데이터 분석에 더 효과가 좋은 3D CNN을 사용하게 되면, 더 좋은 결과를 얻을 수 있을 것으로 보인다.

제3절 수어 데이터 세트

American Sign Language Lexicon Video Dataset (ASSLVD)은 컴퓨터 과학자와 언어학자 간의 협력을 통해 2010년 만들어진 미국의 수어 데이터 세트이다[7]. ASSLVD는 사용자가 미국 수어를 비디오로 녹화하고 컴퓨터 기반의 수어 인식에 의존하여 데이터 세트를 검색할 수 있게 했다. 이를 위해 ASSLVD는 컴퓨터가 수어를 구별하고 인식할 수 있는 알고리즘의 훈련을 위해 단어 당 최대 6명의 수어 사용자로부터 약 3000개의 말뭉치를 만들었다.

<표 2-1> ASSLVD의 통계 개요[7]

수어 종류	수어 개수	변형 수어 개수	사람 당 수어 영상 개수		수어 당 영상		수어 영상 개수
단형 수어	2,284	2,793	x1	621	587	x1	8,585
			x2	989	858	x2	
			x3	394	386	x3	
			x4	563	491	x4	
			x5	85	142	x5	
			x6	141	154	x6	
					175	>6	
복합 수어	289	329	x1	129	117	x1	749
			x2	106	107	x2	
			x3	48	46	x3	
			x4	33	33	x4	
			x5	4	11	x5	
			x6	9	13	x6	
					2	>6	
전체	2,742	3,314	–	–	–	–	9,794

ASSLVD는 <표 2-1>에서 볼 수 있듯이 총 2,742개의 독립된 동작으로 구현된 단어 영상을 수집했으며, 두개 이상의 동작이 복합된 단어의 영상은 별도로 749개의 영상을 수집했다. 이렇게 수집된 영상들은 각각의 단어 당 최대 1~6의 수어 사용자가 촬영을 했으며, 그에 대한 통계는 <표 2-1>에서 볼 수 있다.

ASSLVD은 총4개의 동기화된 카메라를 사용했다. 각각 전신 정면, 전신 측면, 얼굴 부분 확대, 전신 정면(느린 속도)을 촬영했으며, 전신 정면, 전신 측면 및 얼굴은 640 x 480 픽셀의 해상도에 60 프레임의 영상으로 촬영되었고, 느린 속도의 전신 정면 영상은 1600 x 1200의 고해상도에 30프레임의 속도로 촬영됐다. 모든 영상은 수어 사용자들이 미리 준비된 수어 영상들을 보고, 각각의 영상을 자연스럽게 재현하도록 촬영됐다. 하지만 실제 수어의 경우 동일한 단어를 표현함에 있어 수어 사용자 마다 표현방식(예: 손 모양, 손가락 등등)이 미묘하게 다르다. 예를 들어 <그림 2-6>을 보면 동일한 MAN이란 단어를 수어로 표현할 때, 한사람은 시작 시 주먹을 쥐고 있는 반면, 또다른 한사람은 시작 시 손을 펴고 있는 모습을 볼 있다. 이처럼 모든 수어 사용자들은 <그림 2-6>과 같이 준비된 영상과 동일한 동작을 취하진 않았으며, 준비된 영상과 동작이 다를 경우, 이를 해당 영상의 주석에 표시했다.

<그림 2-6> MAN을 표현하는 서로 다른 손 모양

<그림 2-7> ASSLVD의 주석용 손 모양 팔레트의 일부

이러한 주석을 표시하기 위해 ASSLVD는 <그림 2-7>과 같이 데이터 세트에 사용된 모든 손 모양 및 손 모양의 레이블 세트를 만들었다. 각각의 손 모양에 이름과 식별 번호를 할당하여, 수어 영상에 어떤 손 모양을 사용했는지에 대한 주석을 첨부했다. ASSLVD 영상의 주석은 위에서 언급한 손의 모양 뿐만 아닌 수어 사용자의 이름, 단어 명 등을 포함하여 다음 <그림 2-8>와 같은 형태의 엑셀 파일로 제공된다.

B	C	D	E	F	G	H	I	J
ACCIDENT	==========	==========	==========	==========	==========	==========	==========	==========
	Liz	ACCIDENT	ACCIDENT	S	S	S	S	N
	Naomi	ACCIDENT	ACCIDENT	S	S	S	S	N
	----------	----------	----------	----------	----------	----------	----------	----------
	Liz	ACCIDENT	(5)ACCIDENT	5	5	A	A	N
	Naomi	ACCIDENT	(5)ACCIDENT	5	5	10	10	N
	Lana	ACCIDENT	(5)ACCIDENT	5	5	S	S	N
	Dana	ACCIDENT	(5)ACCIDENT	5	5	S	S	N
	----------	----------	----------	----------	----------	----------	----------	----------
	Tyler	ACCIDENT	(3)ACCIDENT	3	3	S	A	N

<그림 2-8> ASSLVD의 주석 일부

Word-Level American Sign Language(WLASL) 데이터 세트는 미국 수어에서 사용되는 2000개의 단어를 사용하는 수어 인식을 위해 2020년에 공개된 가장 큰 비디오 데이터 세트이다. 기존의 미국 수어 데이터 세트는 적은 수의 단어로 제한된다. 이러한 데이터 세트에서 학습한 모델들은 실생활에서 사용하는 단어보다 월등히 적기 때문에 실생활에 적용하기 어렵다.

WLASL에서는 이러한 부분을 개선하고자 100명 이상의 수화 사용자가 수행한 2000개 이상의 단어를 포함하는 새로운 단어 수준 미국 수어 비디오 데이터 세트를 만들어 누구나 사용할 수 있도록 공개하고 있다. 단어 수준의 수어 인식은 수어 문장을 이해하기 위한 가장 기본 구성요소이다. 이 단어를 인식하는 작업 자체도 매우 어려운 과제다.

수어의 단어를 인식하는데 어려운 점은 다음과 같다.

첫째, 수어의 의미는 주로 신체 움직임, 머리 및 손의 위치의 조합에 따라 달라지며, 이는 미묘한 차이로 서로 다른 의미로 해석된다. <그림 2-9>에서 볼 수 있듯 동일한 손모양임에도 불구하고 위치와 손의 방향에 따라 서로 다른 의미를 보여준다.

<그림 2-9> 미국 수어의 “read”(위)와 “dance”(아래)의 손위치 차이

둘째, 동작 인식 및 제스처 인식과 같은 관련 작업에는 최대 수백 개 정도의 범주만 포함이 된다[40, 41, 42, 43]. 그에 비해 매일 사용되는 수어의 어휘는 일반적으로 수천가지 정도가 된다.

셋째, 수어로 표현된 단어는 자연어에서 여러 단어를 가질 수 있다. <그림 2-10>을 보면 동일한 동작이지만 상황에 따라서 다르게 해석이 될 수도 있다. 또한 명사와 동사는 동일한 부호를 가지고 있다.

위와 같은 문제들은 기존의 소규모 데이터 세트에서 잘 포착이 되지 않는다. 실용적인 미국 수어 인식 모델을 만들기 위해서는 충분한 수의 훈련 데이터가 있어야 된다. 이를 위해 WLASL에서는 기존 단어 수준 데이터 세트들이 대규모의 단어 어휘를 제공하지 않는 점을 고려하여 대규모 단어 수준의 수어 영상을 수집했다. 또한 수어 인식에 필요한 최소 하드웨어 요구사항을 활용하기 위해 RGB 기반 비디오 만을 수집했다.

<그림 2-10> 미국 수어의 “wish”(위)와 “hungry”(아래)

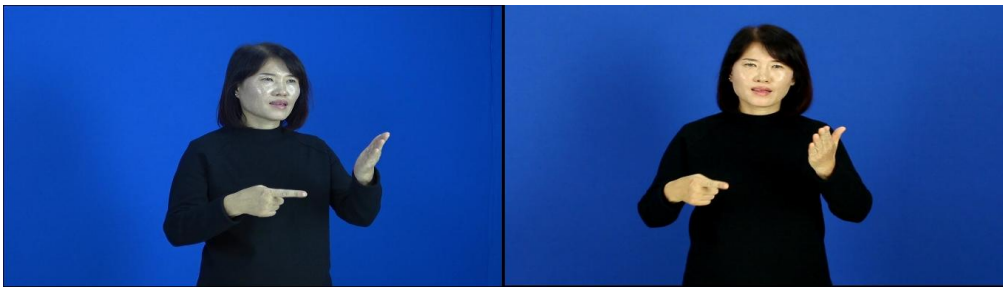
WLASL은 데이터 수집을 위해 인터넷에서 두 가지 주요 영상을 사용한다. 우선 ASLU[43] 및 ASL_LEX[44]와 같은 여러 교육용 수화 웹사이트에서 영상을 수집한다. 교육용 수화 웹사이트의 경우 단어와 수어의 맵핑은 전문가의 확인을 거쳐 업로드가 되기 때문에 정확하다. 두번째 영상은 유튜

브의 미국 수어 튜토리얼 비디오에서 수집되었다. 유튜브의 비디오의 경우 매우 많은 비디오가 있으며, WLASL은 그중 수어와 단어가 명확하게 맵핑이 되는 비디오를 선택하여 수집하였다. 이런 방법으로 총 20개의 서로 다른 웹사이트에서 20,863개의 영단어가 포함된 68,129개의 비디오를 수집한다.

데이터를 수집한 후 단어당 2개의 주석이 있을 경우 한단어만 포함될 수 있도록 해당 영상은 제거한다. 또한 하나의 단어에 대한 비디오가 7개 미만일 경우 훈련용 데이터로 사용하기에는 부족하기 때문에 해당 영상도 제거한다. 마지막으로 대부분의 웹사이트에는 매일 사용되는 단어가 포함되어 있다. 때문에 웹사이트에 해당 단어가 소수만 표출될 경우에도 삭제한다. 이러한 과정을 거치면 3,126개의 단어를 포함한 34,404개의 비디오 샘플이 된다. 그후 각 비디오에 단어 명, 시간 정보(수어의 시작 및 끝 프레임, 반복 여부 등), 몸 영역(YOLOv3 사용), 각 수어 사용자의 ID 및 방언에 대한 주석을 첨부한다. 마지막으로 WLASL은 K = {100, 300, 1000, 2000}인 Top-K 단어를 선택하여 각각 WLASL100, WLASL300, WLASL1000 및 WLASL2000이라는 네가지 데이터 세트로 구성한다. <그림 2-11>는 WLASL 데이터 세트의 일부 구성을 보여준다.

<그림 2-11> WLASL 데이터 세트 예

청각 장애인들이 미처 대처하기 힘들고 어려운 상태에 처해있을 때, 청각장애인들은 외부의 도움을 받는 것이 매우 어렵다. KETI 데이터 세트는 이러한 응급 상황 시 필요한 문장들로 구성되어 있다[14]. KETI 데이터 세트는 105개의 문장과 419개의 단어를 포함하고 있으며, 총 11,578개의 Full HD 해상도 및 초당 30프레임의 비디오로 구성된다. KETI 데이터 세트는 전면과 측면 양방향에서 11명의 청각 장애가 있는 수어 사용자들이 녹화를 했다. 이는 장애인이 아닌 수어사용자의 표현 오류를 제거하기 위함이며, 각 수어 사용자는 데이터 세트에 대해 총 1,048개의 비디오를 녹화했다. <그림 2-12>은 "화재" 단어의 측면 및 정면 촬영 영상의 예를 보여준다.



<그림 2-12> KETI 데이터 세트 중 "화재"의 측면(좌)과 정면(우)의 예

KETI 데이터 세트는 또한 비상사태에서 사용되는 유용한 문장에 해당하는 105개의 문장에는 문장을 구성하는 주요 단어 5개를 선정하여 각각의 영상에 주석으로 첨부했다. <표 2-2>을 보면 "화상을 입었어요."와 같은 문장일 경우 "FIRE" 와 "SCAR"의 단어가 첨부된 것을 볼 수 있으며, "집이 흔들려요."의 경우 "HOUSE"와 "SHAKE"의 단어가 첨부된 것을 볼 수 있다. <표 2-2>의 경우 그림 공간의 문제로 인해 최대 2개씩만 표시했다.

<표 2-2> KETI 수화 데이터 세트 주석의 예[14]

ID	한국어 문장	영어 문장	수어 단어
1	화상을 입었어요.	I got burned.	FIRE SCAR
2	폭탄이 터졌어요.	The bomb went off.	BOMB
3	친구가 숨을 쉬지 않아요.	My friend is not breathing.	FRIEND BREATHE CANT
4	집이 흔들려요.	The house is shaking.	HOUSE SHAKE
5	집에 불이 났어요.	The house is on fire.	HOUSE FIRE
6	가스가 새고 있어요.	Gas is leaking	GAS BROKEN FLOW
7	112에 신고해주세요.	Please call 112.	112 REPORT PLEASE
8	도와주세요.	Help me.	HELP PLEASE
9	너무 아파요.	It hurts so much.	SICK
10	무릎 인대를 다친 것 같아요.	I hurt my knee ligament.	KNEE LIGAMENT SCAR

그러나, KETI의 데이터 세트는 해당 논문[14]의 내용에 설명하고 있는 모든 데이터가 공개되지는 않았다. 1차적으로 공개된 데이터는 11,578개의 비디오가 아닌 8,380개의 비디오였고, 문장에 첨부됐다는 주요 단어 5개의 주석은 찾아볼 수 없었다. 또한 주석에 표시된 파일명은 실제로 존재하는 파일들과 확장자가 상이하다는 문제점이 존재하였다.

제4절 수어 인식 딥러닝 네트워크

앞 절에서 살펴본 바와 같이, 수어 인식은 동작인식 네트워크와는 다르게 공통적으로 사용하는 데이터 세트가 없다. 그 이유는 세계 각 나라마다 사용하는 수어의 체계가 다르기 때문이다. 때문에 대부분의 수어 네트워크

는 해당 국가의 데이터 세트를 사용하게 된다.

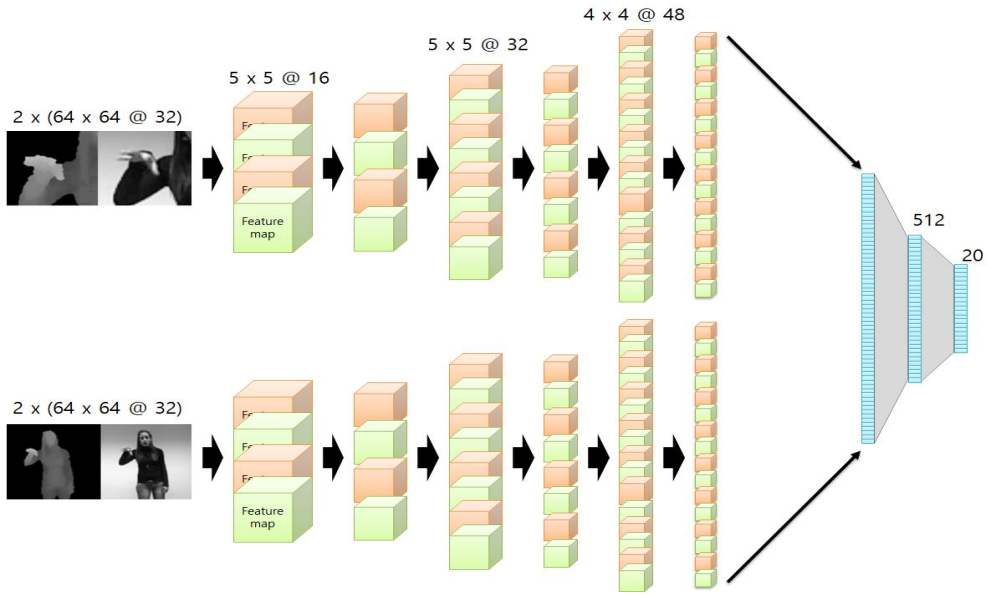
UCF-101 데이터 세트가 아닌 수어 데이터 세트의 경우 매우 다양한 데이터 세트가 존재하기 때문에 네트워크의 성능을 서로 비교하기가 어렵다. L. Pigou[25]의 경우 ChaLearn Looking at People 2014(CLAP14)[51] 데이터 세트를 사용했다. CLAP14는 20종류의 이탈리아어 동작에 대한 데이터 세트로 각 동작은 <그림 2-13>과 같이 RGB, 깊이 영상, 깊이 영상 내 사용자의 위치, 스켈레톤 영상으로 구성되어 있다. L.Pigou 네트워크는 네트워크에 사용하기 위한 입력 영상을 얻기 위해 <그림 2-14>와 같은 전처리 과정을 수행한다. 전처리 과정의 첫번째 과정은 손과 상체 영상을 자르는 것이다. CLAP14의 이탈리어 동작은 양손은 동일한 움직임을 수행한다. 따라서 네트워크 모델은 양손 중 한쪽 손만 훈련하면 된다. 또한 손은 항상 상체 주위에 있기 때문에 네트워크 훈련에 필요한 영상은 한 손의 영상과 상체의 영상만 있으면 된다. 그 다음 과정은 깊이 맵에서 잡음을 줄이는 과정이다. 임계값을 사용해 노이즈를 줄이고, User Index를 사용해 배경을 지운다. 이러한 과정을 통해 64 x 64크기에 32프레임으로 구성된 4개의 입력 비디오 영상(그레이 스케일의 손, 상체 및 깊이 맵의 손,상체)을 얻는다.

<그림 2-13> CLAP14의 데이터 세트 구성

(a) Original (b) Noise reduction (c) 4 Input channels

<그림 2-14> L. Pigou 네트워크의 전처리 과정

L. Pigou의 네트워크는 데이터 셋의 영상을 전처리 과정을 통해 상체와 손부분으로 잘라내 사용하였으며, <그림 2-15>와 같이 2개의 CNN 네트워크로 구성되어 있다. 네트워크 중 하나는 손 특징을 위한 네트워크이며, 또 다른 하나는 상체의 특징을 추출하기 위한 네트워크다. 네트워크에 사용된 풀링은 3D 최대 풀링이 사용되었으며, 컨볼루션 레이어는 2D CNN인 Alexnet[28]을 사용했다. 손과 상체를 학습하는 각 네트워크는 하나의 커볼루션 레이어와 하나의 최대 풀링으로 구성된 3개의 컨볼루션 그룹으로 구성된다. 3개의 컨볼루션 그룹을 지난 각각의 네트워크를 지난 데이터는 하나로 합쳐 완전 연결 레이어에 전달된다. 가장 처음 컨볼루션 레이어와 두번째 컨볼루션 레이어에는 지역 콘트라스트 정규화(LCN)[58]이 적용되었으며, 모든 컨볼루션 레이어의 활성화 함수로는 ReLU[59, 60]를 사용했다.

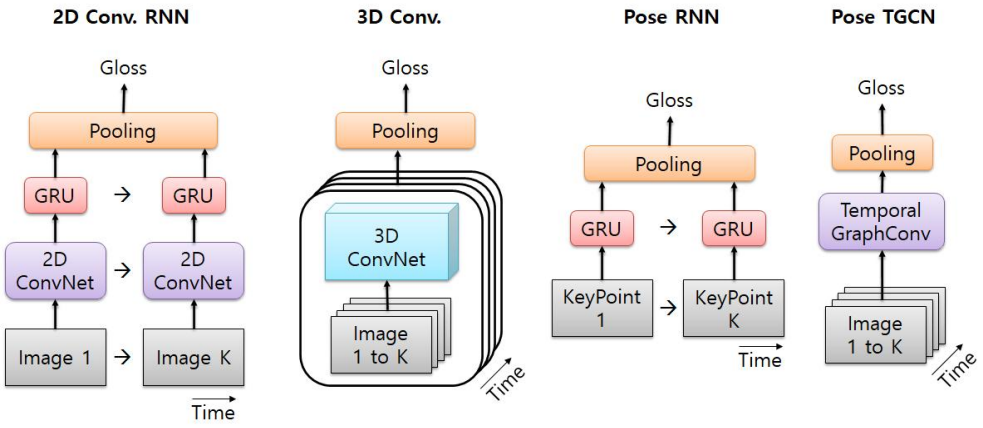


<그림 2-15> L. Pigou 네트워크의 구성도

<표 2-3> L. Pigou 네트워크 검증 결과

	에러율(%)	개선률(%)
Tanh units	18.9	
ReLU	14.4	23.8
+ dropout	11.9	17.4
+ LCN(처음 2레이어)	10.3	13.4
+ data augmentation	8.3	19.4

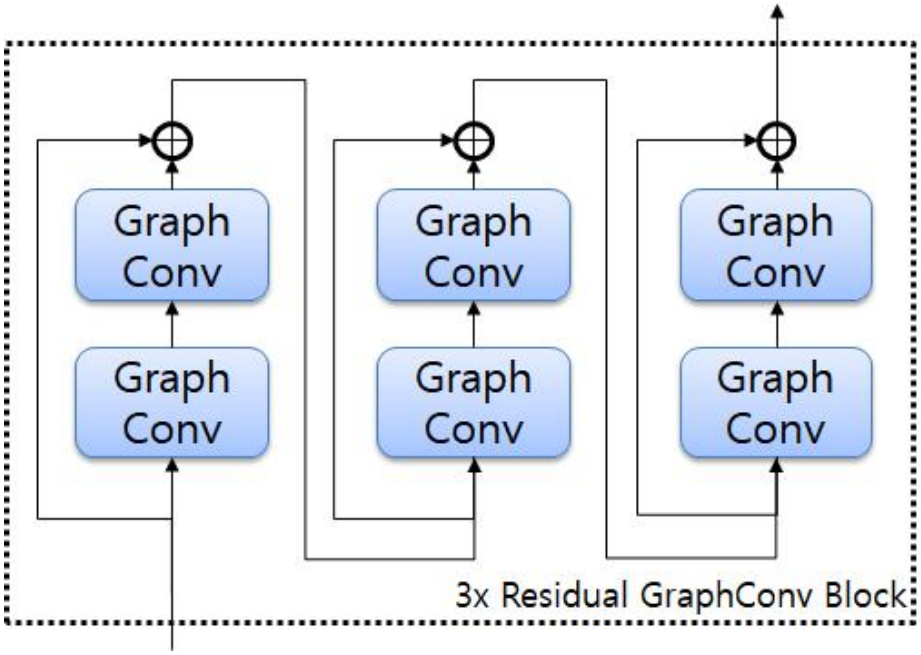
L. Pigou 네트워크는 학습률 0.003에 감소율 5%로 설정 하여 네트워크를 학습했으며, 그 결과는 <표 2-3>과 같다. 활성화 함수로 Tanh를 사용했을 때와 비교해서 전체적으로 23.8%의 정확도 개선이 이루어 졌으며 최종 구성에서는 91.7%의 정확도, 8.3%의 오차율을 보인다.



<그림 2-16> D.Li에서 사용된 네트워크의 구성도

D. Li[37]의 경우는 기존 WLASL 데이터 세트를 새로 만들었으며, 4개의 네트워크를 사용해 데이터 세트를 평가 했다. 사용한 4개의 네트워크는 <그림 2-16>와 같이 CNN+GRU, 3D CNN, Pose RNN 및 Pose TGCN이다. CNN+GRU와 3D CNN은 이미지를 기반으로 하는 딥러닝 네트워크다. GRU[60]는 LSTM의 한 변형 모듈로 LSTM과 구조상 큰 차이가 없으며, 성능 또한 비슷하다. 하지만 GUR는 LSTM보다 학습할 가중치가 적다는 이점을 가진다. D.Li의 실험에서 사용된 CNN+GRU은 ImageNet으로 훈련된 VGG16 네트워크를 사용해 공간적 특징을 추출한다. 그후 64, 96, 128, 256개로 구성된 GRU를 지나 평균 풀링을 거쳐 결과를 내보낸다. 입력 영상으로는 동영상당 무작위 위치에서 뽑은 연속된 50 프레임을 사용한다. D.Li의 실험에 사용된 3D CNN은 I3D[43] 네트워크를 사용한다. ImageNet으로 미리 학습되어 있으며, Kinetics-400[43]으로 파인 튜닝된 모델이다. D.Li의 실험에서는 I3D 모델의 마지막 분류 레이어만 WLASL의

클래스 개수에 맞도록 수정하여 사용했다. Pose RNN과 Pose TGCN은 사람의 자세에 기반한 딥러닝 네트워크다. 사람의 자세 추정은 단일 이미지 혹은 비디오에서 인체의 키포인트 또는 관절을 찾아 특정하는 것을 목표로 한다[52, 61, 62, 63, 64, 65, 66, 67]. D.Li의 실험의 Pose RNN은 OpenPose를 사용해서 인체의 키포인트를 추출한다. 55개의 몸 키포인트를 추출하고, 64, 64, 128, 128로 구성된 GRU에 전달한다. 입력 영상은 각 영영상으부터 무작위 위치의 연속된 50프레임을 추출하여 사용한다. Pose TGCN은 D.Li에서 제안하는 통계학습에 대한 새로운 자세 기반 접근 방식이다. TGCN은 자세 시퀀스의 공간적 및 시간적 의존성을 모델링하는 새로운 그래프 네트워크 기반 구조를 가진다. 일반적으로 2D 관절 각도를 사용하여 움직임을 모델링하는 사람의 자세 추정에 대한 기존 연구와는 달리, 시간적 움직임 정보를 신체 키 포인트의 궤적에 대한 전반적인 내용으로 표현한다. 인간의 자세 예측에 관한 연구[68]에 의해 D.Li의 연구에서는 인체를 하나의 꼭지점으로부터 완전히 연결된 그래프로 본다. 인체는 부분적으로 연결되어 있지만 그래프 네트워크를 통해 각 관절간의 종속성을 훈련하기 위해 인체를 완전히 연결된 그래프로 재구성한다. TGCN은 재구성된 데이터를 입력으로 사용하고, <그림 2-17>의 RGC 블록을 통해 처리된다. RGC 블록은 2개의 그래프 컨볼루션 레이어를 잔차 연결로 구성한다.



<그림 2-17> TGCN의 RGC 블록

TGCN은 다중 잔차 그래프 컨볼루션 블록을 쌓아두고, 사람의 자세 궤적 특징 표현으로 시간적 특성을 취한다. 그후 소프트 맥스 레이어와 평균 풀링 레이어를 사용해 데이터를 분류한다. D.Li에서 상기 4개의 네트워크를 이용한 실험 결과는 다음 <표 2-4>와 같다.

<표 2-4> D.Li 네트워크 실험 결과

	WLASL100	WLASL300	WLASL1000	WLASL2000
Pose GRU	46.51	33.68	30.01	22.54
Pose TGCN	55.43	38.32	34.86	23.65
CNN GRU	25.97	19.31	14.66	8.44
I3D	**65.89**	**56.14**	**47.33**	**32.48**

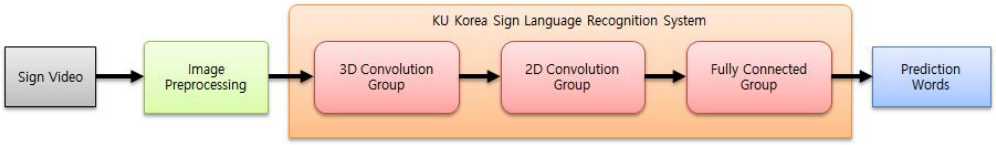
4개의 네트워크중 3D CNN을 사용한 I3D가 가장 뛰어난 성능을 보여주며, 그 뒤를 D.Li에서 제안한 TGCN이 따르고 있다. WLASL100에서는 정확도 65.89%, WLASL300은 56.14%, WLASL1000은 47.33%, 마지막으로 WLASL2000은 32.48%의 정확도를 보여준다.

KETI[14]는 한국 수어 데이터 세트인 KETI 데이터로부터 OpenPose[52, 53, 54]를 사용하여 스켈레톤 영상을 구하고 다시 특징소를 추출하여 사용하였다. OpenPose는 영상의 사람에 대한 스켈레톤 영상을 추출한다. 추출된 사람의 스켈레톤 영상은 124개의 키 포인트로 구성되며, 몸 12포인트, 각 손당 21포인트, 얼굴 70포인트로 구성되어 있다. KETI는 스켈레톤 영상 추출 후 총 124개의 포인트 중 얼굴포인트는 사용하지 않는다. 얼굴 포인트를 제외한 54개 키포인트 값들을 대상으로 LSTM 네트워크를 이용해 수어를 인식한다. 그 결과 단어 인식의 정확도는 55.28%다.

본 절에서 살펴본 바와 같이 한국 수어는 그 데이터 양도 적고 신뢰하기 힘든 데이터 셋이 많다. 또한 단어의 인식 정확도 마저 매우 낮은 현실이다. 이를 해결하기 위해 본 논문에서는 시계열 데이터를 처리하는데 뛰어난 3D CNN과 OpenPose의 스켈레톤 영상을 사용하여 한국 수어 인식의 정확도를 높이고자 한다.

제3장 한국 수어 단어 인식 네트워크 설계 및 구현

제1절 KU 한국 수어 인식 시스템 구조

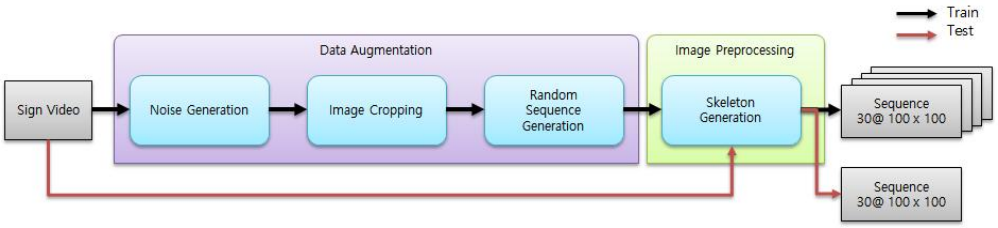


<그림 3-1> KU 한국 수어 인식 시스템 구조도

본 장에서는 한국 수어 자동 인식을 위한 심층 신경망 네트워크를 설계하고, 네트워크의 성능 평가를 위한 데이터 세트를 제안한다. 제안하는 KU 한국 수어 인식 시스템은 <그림 3-1>와 같다. KU 한국 수어 인식 시스템은 수어 영상을 입력으로 사용한다. 입력된 영상은 이미지 전처리 과정을 통해 스켈레톤 영상으로 만들어진 30 프레임 길이를 갖는 100 x 100 해상도의 시퀀스로 변환된다. 변환된 시퀀스는 3D 컨볼루션 레이어, 2D 컨볼루션 레이어 및 완전 연결 레이어로 구성된 네트워크를 통해 수어 영상의 단어를 예측한다.

KU 한국 수어 인식 시스템은 3D 컨볼루션 레이어를 통해 시공간에 대한 특징들을 추출하고, 2D 컨볼루션을 적용하여 한번 더 공간적 특징에 대한 특징을 추출한다. 추출된 특징은 완전 연결 레이어로 전달되고, 완전 연결 레이어를 통해 수어 단어를 예측한다

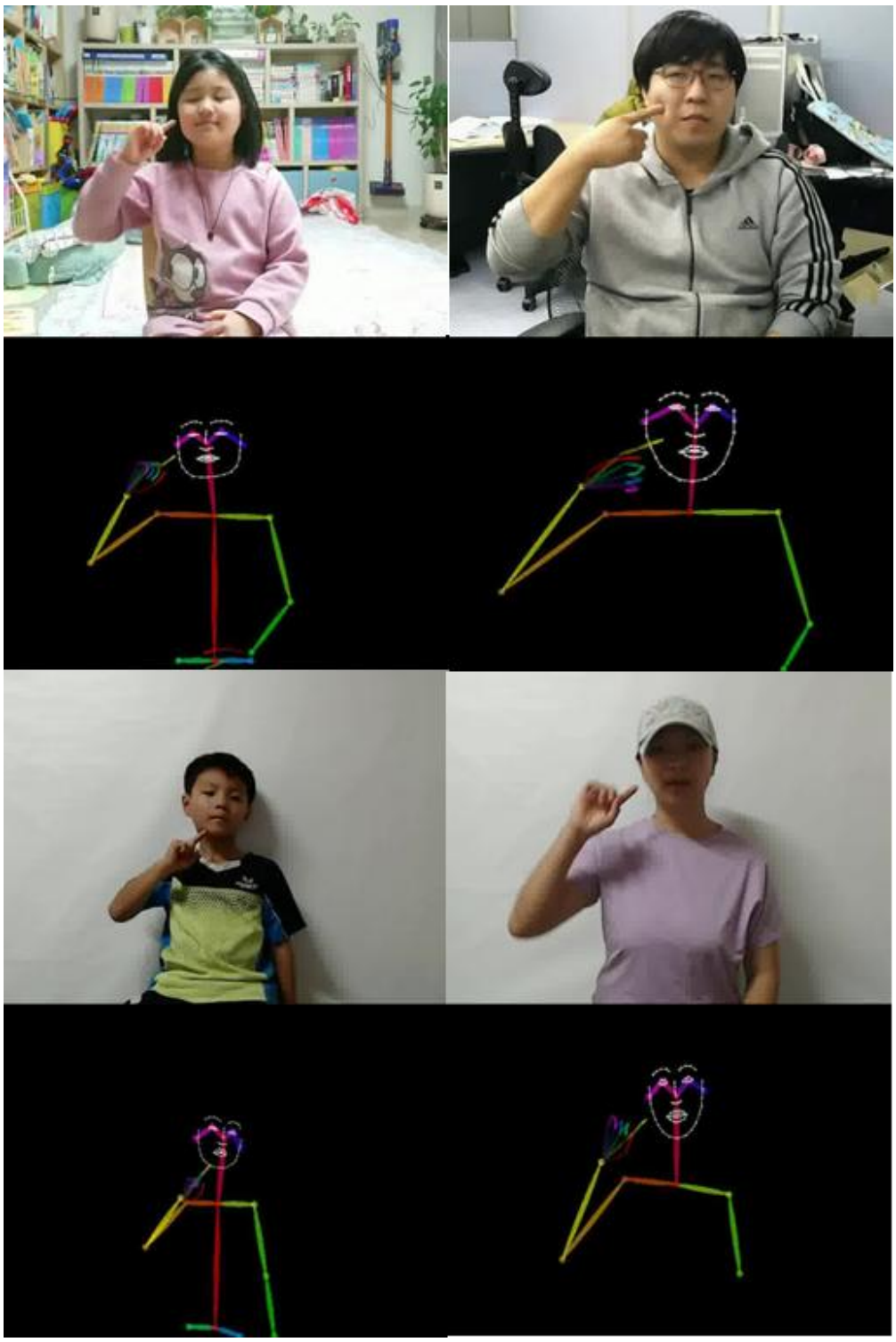
제2절 KU 한국 수어 인식 시스템을 위한 데이터 전처리



<그림 3-2> KU 한국 수어 인식 시스템 전처리 과정

KU 한국 수어 인식 시스템을 위한 전처리 과정은 KU 한국 수어 인식 시스템에서 수어 인식률을 높이기 위한 중요한 과정이다. KU 한국 수어 인식 시스템을 위한 데이터 전처리 과정은 <그림 3-2>에서 보이는 것과 같이 네트워크의 훈련 및 실험에서 모두 수행되는 과정이다.

영상 처리에서 일반적으로 사용되는 RGB 영상은 카메라를 통해 얻을 수 있는 가장 기본적인 영상 데이터다. 하지만 이런 RGB 영상은 상황에 따라 영상내 잡음이 발생하며, 배경으로 인하여 원하는 정보의 추출이 매우 어렵다. 때문에 본 논문에서는 영상내 잡음과 무관한 영상을 추출하기 위하여, 기존 라이브러리인 OpenPose를 사용해 RGB영상으로부터 사람의 스켈레톤을 추출하여 사용하는 방법을 제안한다. OpenPose는 다양한 환경에서 매우 안정적으로 스켈레톤 영상을 추출할 수 있다. 또한 사람의 얼굴 표정 및 손가락까지도 매우 정교하게 추출한다. 수어에서는 사람의 손 동작뿐 만 아닌 얼굴 표정, 손가락의 모양까지도 모두 중요한 의미를 갖는다. 때문에 OpenPose는 사람의 스켈레톤 영상을 추출하기 위한 매우 좋은 선택이다. <그림 3-3>는 OpenPose로 추출한 스켈레톤 영상의 일부를 보여준다.



<그림 3-3> 수어 동영상 스켈레톤 추출 예

제3절 KU 한국 수어 단어 인식 네트워크

본 절에서는 한국 수어를 높은 정확도로 인식하기 위한 단어 인식 딥러닝 네트워크 모델을 구성하였다. 단어는 문장을 구성하는 기본 요소이며, 이는 수어의 문장을 구성하는 기본요소가 된다. 하지만 한국 수어의 단어 인식 연구는 매우 미흡한 상황이다. 본 논문에서는 한국 수어의 낮은 단어 인식률을 해결하기위해 여러 딥러닝 모델을 구성하고 실험 했다. <표 3-1>과 <표 3-2>는 본 논문에서 실험에 사용한 모델들의 구조를 보여준다.

첫번째 모델은 LRCN 모델로, 2D-CNN 과 LSTM을 같이 사용하는 모델이다. 2D-CNN은 InceptionV3 모델을 사용하였으며, ImageNet 가중치를 사용하여 특징을 추출한다. 그 후 LSTM 레이어 2개를 거쳐 완전 연결 레이어로 전달되어 결과를 도출한다. 두번째 모델은 Merge 3D CNN 모델이다. 하나의 영상으로부터 서로 다른 시퀀스 2개를 추출하여 입력 영상으로 사용한다. 각 입력 영상은 3개의 3D 컨볼루션 레이어와 최대 풀링 레이어를 지나 병합되고, 병합된 특징들은 완전 연결 레이어로 전달된다. 세번째 모델은 C3D[36] 모델이다. C3D 모델은 동작 인식 연구에서 많이 사용되는 모델이며, 높은 정확도를 보여주는 모델이다. 네번째 모델은 Compact C3D 모델로 기존 C3D 모델에서 마지막 컨볼루션 영역을 제외한 모델이다. 또한 완전 연결 레이어의 구성을 4096개에서 1024개로 변경하여 3D 컨볼루션 영역에서 완전 연결 영역으로 넘어가는 가중치에 대한 개수를 증가시켰다. C3D모델과 비교 시, 완전연결 영역에 전달되는 특징 개수 및 가중치 수에 대한 상관관계를 확인할 수 있다. 다섯 번째 모델은 Simple 3D CNN 모델이다. Simple 3D CNN 모델은 4개의 컨볼루션 레이어, 3개의 최대 풀링 레이어 및 3개의 완전연결 레이어로 구성되어 있다. 다른 모델에 비해 적은

필터의 개수와 낮은 깊이를 지닌 모델이다. 여섯 번째 모델은 Middle 3D CNN 모델이다. 4개의 컨볼루션 레이어, 4개의 최대 풀링 레이어 및 2개의 완전 연결 레이어로 구성되어 있으며, 각 컨볼루션 레이어의 필터 개수는 Simple 3D CNN보다 더 많은 레이어다. 대신 완전 연결 레이어를 하나 감소시켜, 동일한 깊이의 모델일 경우 컨볼루션 영역에 치중한 경우와 완전 연결 레이어에 치중한 경우에 따른 결과를 확인해 볼 수 있다. 마지막으로는 KU 수어 단어 인식 네트워크 모델이다. KU 수어 단어 인식 네트워크 모델은 본 논문에서 제안하는 모델이며, C3D[36]의 구조를 참조하여 확장한 모델이다. 일반적으로 딥 러닝 네트워크의 경우 입력영상에 따라 네트워크 구성이 변화하게 된다. KU 수어 단어 인식 네트워크는 100 x 100의 해상도와 30 프레임의 길이를 가지는 입력영상에 맞춰 네트워크를 수정 및 최적화가 되었다. KU 수어 단어 인식 네트워크 모델과 C3D의 가장 큰 차이는 입력 데이터의 크기 및 추가 컨볼루션 레이어 영역이다. 입력 데이터의 크기를 특정하여 데이터의 정확도를 높였으며, 완전 연결 레이어로 진행되기 전에 2D 컨볼루션 레이어를 추가하여 정확도를 끌어올렸다. C3D에 있는 제로 패딩 레이어를 제거하였으며, 대신 컨볼루션 레이어를 추가해 정확도를 향상시키고자 했다.

<표 3-1> 네트워크 실험에 사용된 네트워크 구조 1

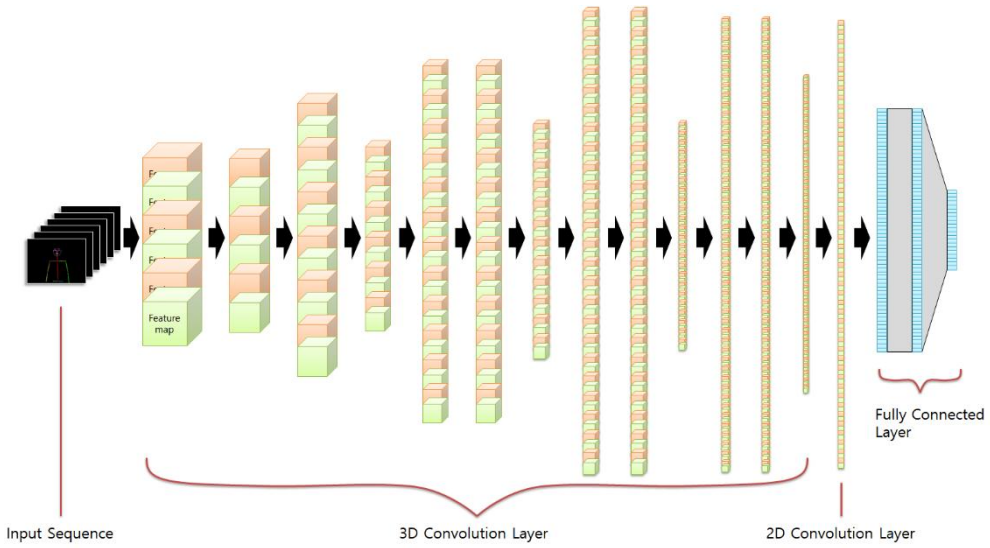
	LRCN	Merge 3D CNN		C3D
1	InceptionV3	Conv 32,5,5,5	Conv 32,5,5,5	Conv 64,3,3,3
2	LSTM 2048	Pooling	Pooling	Pooling
3	LSTM 2048	Conv 64,5,5,5	Conv 64,5,5,5	Conv 128,3,3,3
4	FC 2048	Pooling	Pooling	Pooling
5	FC 2048	Conv 128,3,3,3	Conv 128,3,3,3	Conv 256,3,3,3
6	FC 1024	Pooling	Pooling	Conv 256,3,3,3
7	FC 41	Concatenate		Pooling
8		FC 1024		Conv 512,3,3,3
9		FC 512		Conv 512,3,3,3
10		FC 41		Pooling
11				Conv 512,3,3,3
12				Conv 512,3,3,3
13				Pooling
14				ZeroPadding
15				Pooling
16				FC 4096
17				FC 4096
18				FC 41

<표 3-2> 네트워크 실험에 사용된 네트워크 구조 2

	Compact C3D	Simple 3D CNN	Middle 3D CNN	KU 수어 단어 인식 네트워크
1	Conv 64,3,3,3	Conv 32,3,3,3	Conv 64,3,3,3	Conv 64,3,3,3
2	Pooling	Pooling	Pooling	Pooling
3	Conv 128,3,3,3	Conv* 64,3,3,3	Conv 128,3,3,3	Conv 128,3,3,3
4	Pooling	Pooling	Pooling	Pooling
5	Conv 256,3,3,3	Conv 128,3,3,3	Conv 256,3,3,3	Conv 256,3,3,3
6	Conv 256,3,3,3	Conv 128,3,3,3	Pooling	Conv 256,3,3,3
7	Pooling	Pooling	Conv 512,3,3,3	Pooling
8	Conv 512,3,3,3	FC 1024	Pooling	Conv 512,3,3,3
9	Conv 512,3,3,3	FC 1024	FC 1024	Conv 512,3,3,3
10	Pooling	FC 41	FC 41	Pooling
11	FC 1024			Conv 512,3,3,3
12	FC 1024			Conv 512,3,3,3
13	FC 41			Pooling
14				Reshape
15				Conv2D 1024,2,2
16				Pooling2D
17				FC 4096
18				FC 4096
19				FC 41

<표 3-1>와 <표 3-2>의 Conv는 모두 3D 컨볼루션 레이어를 뜻하며, Pooling은 3차원 최대 값 풀링을 의미한다. 2D 컨볼루션의 경우 Conv2D로 표기했으며, 2D 풀링의 경우에는 Pooling2D로 언급했다.

KU 수어 단어 인식 네트워크의 입력 데이터는 OpenPose에서 추출한 스켈레톤 영상으로 크기는 100x100이며, 30개의 프레임으로 구성된 시퀀스를 사용한다. KU 수어 단어 인식 네트워크는 <그림 3-4>와 같이 크게 3D 컨볼루션 레이어 영역과 2D 컨볼루션 레이어 영역 및 완전연결 레이어 영역으로 구성되며, 입력 영상의 크기인 100x100의 크기와 시퀀스의 길이인 30프레임에 맞춰 구성되어 있다.



<그림 3-4> KU 수어 단어 인식 네트워크 구성도

여기서 3D 컨볼루션 레이어 영역은 컨볼루션 레이어1, 풀링 레이어1, 컨볼루션 레이어2, 풀링 레이어2, 컨볼루션 레이어3, 컨볼루션 레이어4, 풀링 레이어3, 컨볼루션 레이어4, 컨볼루션 레이어5, 풀링 레이어4, 컨볼루션

레이어6, 컨볼루션 레이어7, 풀링 레이어5 등 모두 7개의 컨볼루션 레이어와 5개의 풀링 레이어가 교차로 구성된다. 3D 컨볼루션 레이어 영역의 모든 컨볼루션 레이어는 3D 컨볼루션을 사용하고 스트라이드 값은 1이다. 각 컨볼루션 레이어의 필터 개수는 64, 128, 256, 256, 512, 512, 512, 512 이며, 모든 필터의 크기는 3 x 3 x 3이다. 모든 풀링 레이어는 최대 풀링을 사용하며, 풀링 레이어1를 제외한 모든 레이어의 풀링 크기는 2 × 2 × 2, 스트라이드는 2 × 2 × 2의 값을 가진다. 풀링 레이어1은 1 × 2 × 2의 크기를 가진다. 3D 컨볼루션 레이어 영역이 수행되면 데이터는 1 x 3 x 3 x 512의 데이터 형태가 된다. 이때 가장 앞에 있는 1을 제거해도 데이터에는 아무런 손실이 발생하지 않는다. 따라서 해당 데이터를 3 x 3 x 512의 형태로 변경하고 2D 컨볼루션을 한 번 더 진행한다. 2D 컨볼루션은 기본적으로 3D 컨볼루션보다 자원을 적게 소모하며, 전체 네트워크에 추가 레이어의 첨가로 인해 정확도는 올라간다. 마지막으로 4096개로 구성된 완전 연결 레이어 2개를 지나 단어를 예측하게 된다.

제4절 KU 한국 수어 데이터 세트

1. 데이터셋 수집

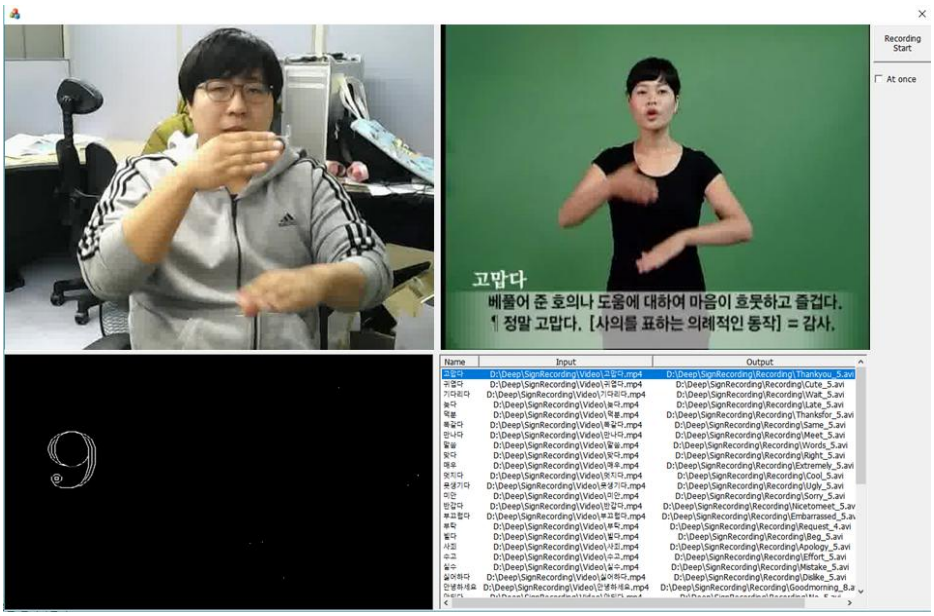
KU 한국 수어 데이터 세트는 국립국어원의 한국수어사전의 카테고리를 참조한 일상 생활에서 많이 사용되는 인사 및 대화를 주제로 삼는 41개의 단어를 선택하여 구성했다. 41개의 단어는 최소 27개, 최대 31개의 영상을 가지고 있으며, 각 영상은 1280 x 720 @30FPS로 정면을 녹화했다. 총 영상의 수는 1151개, 총 프레임은 131,983 개이다. 10명의 일반인이 <그림 3-5>와 같은 한국 수어 사전에서 제공하는 영상을 시청하고 동일한 동작

으로 촬영했다. 촬영은 실생활에서 접할 수 있는 다양한 배경을 가지고 촬영했으며, 동일한 사람이 재촬영 할 경우에는 <그림 3-6>과 같이 복장 및 액세서리를 변경하고 촬영을 진행했다. 또한 상황의 다양성을 위해 카메라와 촬영자의 거리를 변경하면서 촬영을 진행했다.

<그림 3-5> 국립국어원 한국수어사전의 일부영상

<그림 3-6> KU 한국 수어 데이터 세트의 동일인 촬영 예

원활한 촬영을 진행하기 위해 <그림 3-7>의 수어 데이터 세트 촬영 프로그램을 제작하여 촬영을 진행했다. 수어 데이터 세트 촬영 프로그램은 촬영을 시작하면, 우측 상단의 샘플영상이 3초 먼저 실행되고, 그후 촬영을 시작한다. 해당 단어의 수어 동작이 끝나면 자동으로 다음 단어로 진행하여 41개의 단어에 대한 수어 동작을 모두 촬영한다. 촬영되는 단어의 목록은 아래 <표 3-3>과 같다.



<그림 3-7> 수어 데이터 세트 촬영 프로그램

KU 한국 수어 데이터의 모든 영상은 <표 3-4>와 같이, 영상에 해당하는 수어 단어 이름, 재생시간, 영상의 총 프레임 수, 파일명과 같은 주석이 영상의 촬영 시 첨부된다.

<표 3-3> KU 한국 수어 데이터 세트 단어 목록

사죄	빌다	조심	연락	멋지다	귀엽다	싫어하다	수고
부끄럽다	매우	앞으로	즐겁다	주십시오	좋다	안녕하세요	친절
늦다	좋아하다	만나다	실수	절대로	반갑다	안되다	의견
예쁘다	질문	정말	화해	안타깝다	부탁	맞다	똑같다
미안	죄송하다	덕분	고맙다	못생기다	용기	기다리다	천만에
말씀							

<표 3-4> KU 한국 수어 데이터 세트 주석 예

수어 이름	재생 시간	프레임 수	파일명	영문명
사죄	00:06	172	Apology_10.avi	Apology
사죄	00:04	131	Apology_21.avi	Apology
예쁘다	00:04	106	Pretty_40.avi	Pretty
빌다	00:03	90	Beg_10.avi	Beg
기다리다	00:06	172	Wait_22.avi	Wait

딥러닝 네트워크를 이용해 데이터를 훈련하는데 가장 중요한 것은 역시 데이터의 수다. 단순히 데이터가 많다고 훈련이 잘되는 것은 아니며, 훈련을 하고자 하는 각각의 클래스들의 데이터의 수가 중요하다. KU 한국 수어 데이터세트와 다른 한국 수어 데이터 세트들의 비교를 보여주는 <표 3-5>에서 단어 당 영상의 수가 바로 각 클래스들의 데이터 수를 의미한다. KETI

데이터 세트는 단어 당 영상의 수가 20개로 양호한 수량을 보여준다. 하지만 KETI 데이터 세트는 정면과 측면 두개의 영상을 포함하고 있다. 이는 측면의 영상의 경우 실제로 수어를 카메라 앞에서 사용할 때는 카메라 정면에서 사용하는 경우가 많다는 걸 고려하면, 도움이 되지 않는 영상이라 할 수 있다. 따라서 KETI 데이터 세트의 측면 영상을 제외하게 되면 KETI 데이터 세트는 단어 당 영상의 수는 10개로 줄어들게 된다. 반면 KU 한국 수어 데이터 세트는 포함하고 있는 단어의 수는 적지만, 단어 당 영상의 수가 평균 28개로 가장 많다. 이는 <표 3-6>의 전 세계 다른 수어 데이터 세트와 비교해도 가장 많은 수를 보여준다.

<표 3-5> 다른 한국 수어 데이터 세트와의 비교

데이터세트	단어	촬영자	총 프레임	총 영상	단어 당 영상 수
KETI[14]	415	14	3,022,600	14,672	20개(정면,측면)
KSL[3]	77	20	112,564	1,229	평균 16개
KU	41	10	131,983	1,151	평균 28개

<표 3-6> 세계 다른 수어 데이터 세트와의 비교

데이터세트	단어	촬영자	총 영상	단어 당 영상
Purdue RVL-SLL[6]	39	14	546	14
RWTH-BOSTON-50[55]	50	3	483	3
Boston ASSLLVD[7]	2,742	6	9,794	6
WLASL100[37]	100	97	2,038	20
WLASL2000[37]	2000	119	21,083	10
KU	41	10	1,151	28

2. 데이터 증강

KU 한국 수어 데이터 세트는 다른 수어 데이터 세트들에 비해 단어 당 영상의 수가 많지만, 여전히 딥러닝 네트워크에 훈련용으로 사용하기에는 수량이 부족한 점이 있다. 그것은 부족한 데이터 세트로 딥러닝 네트워크 모델을 훈련하게 되면, 훈련된 모델은 일반화가 되지 않기 때문이다. [46]의 경우는 영상의 일부 영역을 따와서 하나의 데이터로 만들어 데이터를 보강한다. 또한 [14]의 경우, 동영상의 프레임을 랜덤으로 추출하여 새로운 시퀀스를 만들어 데이터를 보강한다.

본 논문에서는 부족한 데이터를 딥러닝 네트워크 모델에 적합한 데이터로 만들기 위해 다음과 같은 3가지 데이터 보강 방법을 사용한다.

- 동영상 자르기
- 영상 내 잡음 생성
- 무작위 프레임 추출

동영상 자르기는 <그림 3-8>과 같이 영상의 좌, 우, 위, 아래 및 가운데의 영역을 설정 후 영상을 자른다. 실제 생활에서 수어 사용자가 카메라 앞에 있을 때 정확히 가운데 위치하기는 쉽지 않은 일이다. 상황에 따라서는 좌측으로, 또는 우측으로 치우칠 수도 있으며, 카메라의 위치 및 수어 사용자의 위치에 따라 위 혹은 아래쪽으로 치우치게 될 수도 있다. 본 논문에서 사용한 동영상 자르기는 이러한 상황들에 좀더 강인한 데이터를 만들기 위한 방법이다.

<그림 3-8> 동영상 자르기 영역

다음 방법은 영상에 잡음을 추가하는 방법이다. 수어 사용자가 카메라 앞에서 수어 동작을 취할 때 주변의 광원은 항상 완벽한 상태가 아니다. 또한 카메라의 성능 및 노후화로 인해 촬영된 영상에 잡음이 생길 수 있다. 영상에 잡음을 추가하여 훈련 데이터에 포함시키는 이 방법은 위와 같은 상황에서 보다 강인한 데이터가 된다. 본 논문에서 데이터 세트에 적용한 잡음은 가우시안 잡음이다. 가우시안 잡음은 가우시안 모델을 따르는 잡음으로 가우시안 분포의 평균값과 표준 편차에 따라 영상에 잡음을 생성한다. KU 수어 데이터 세트에는 <그림 3-9>와 같이 평균 15, 표준편차 15인 가우시안 잡음을 추가한 영상을 생성했다.

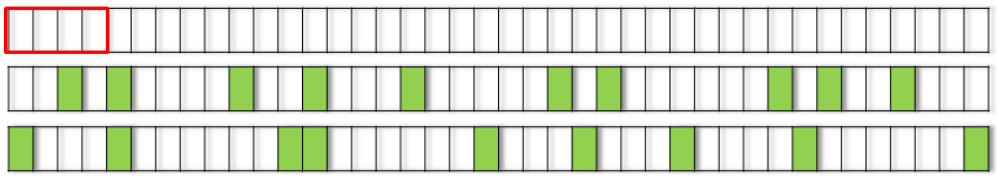
<그림 3-9> 가우시안 잡음 적용 영상

좌측 위: 평균(0), 표준편차(0), 우측 위: 평균(0), 표준편차(15)

좌측 아래: 평균(15), 표준편차(0), 우측 아래: 평균(15), 표준편차(15)

마지막 데이터 보강 방법은 무작위 프레임 추출 방법이다. 딥러닝 네트워크 모델에서 동영상을 훈련할 때 모든 프레임을 데이터로 사용하게 되면 결과는 향상될 것으로 예상된다. 하지만 동영상을 훈련하는 것은 매우 많은 하드웨어 자원을 필요로 한다. 또한 입력영상의 길이는 단어 별로 달라 질 수 있으며, 이 경우 네트워크 훈련 및 시험 시 입력영상의 길이가 달라지기 때문에 동영상을 처리하는 네트워크는 해당 동영상으로부터 정해진 일정한 개수의 프레임을 추출 후 시퀀스를 만들어 처리하게 된다. 추출하는 방법은

다양하다. 무작위로 연속된 프레임을 추출하는 방법[37, 43, 46], 영상의 프레임을 특정 크기로 분할 후 분할한 프레임들 내에서 무작위로 추출하는 방법[14, 57]등이 있다. <그림 3-10>은 전체 프레임을 균등하게 분할 후 분할된 영역내에서 임의로 프레임을 하나씩 추출하여 시퀀스를 만드는 예를 보여준다. 가장 상단의 빨간색 네모 영역처럼 프레임을 분할하고, 그 분할한 영역내에서 무작위로 하나의 프레임을 추출한다. 이렇게 각 영역에서 하나씩 무작위로 프레임을 추출하여 전체 시퀀스를 만든다. 하지만 이 방법은 시퀀스를 구성하는 프레임의 길이가 길면 길수록 무작위성을 보장하기 힘들다. 그 이유는 수어 단어를 구성하는 프레임은 약 115개의 프레임이다. 시퀀스를 구성하는 프레임을 20개의 프레임으로 설정하면 각 프레임과 프레임의 차이는 약 5프레임의 차이가 난다. 동영상에서 5프레임은 움직임에 큰 차이를 보이지 않는다.



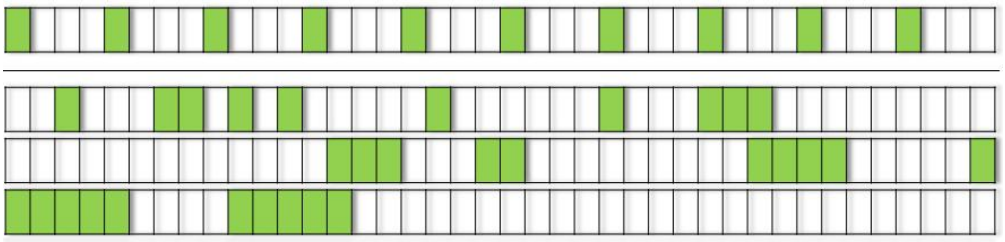
<그림 3-10> 영상 분할 후 무작위 추출의 예

본 논문에서는 <그림 3-11>과 같이 두가지 방법을 사용해 영상 시퀀스를 만든다. 첫째, 영상에서 일정한 간격으로 프레임을 추출해서 시퀀스를 만든다. 추출된 시퀀스를 S라 하고, 해당 영상의 전체 프레임 수를 N이라 한다. 시퀀스를 구성하는 프레임의 수는 K라 할 때 식 (3-1)과 같이 계산하여 시퀀스를 하나 구한다.

$$S = \left\{\frac{N}{K}\times 1, \frac{N}{K}\times 2, \frac{N}{K}\times 3 \dots\dots\dots, \frac{N}{K}\times K\right\} \quad (3-1)$$

또다른 방법은 영상의 전체 프레임 중 무작위로 K개를 추출하여 시퀀스S를 2개 만든다. 기존의 프레임을 분할 후 분할한 프레임에서 추출하는 방법과는 다르다.

본논문에서는 영상을 프레임단위로 분할하지 않고 <그림 3-11>의 아래와 같이 전체 프레임을 기준으로 무작위로 K개의 프레임을 추출한다. 이는 영상의 앞, 중간 및 뒷부분 등 다양한 방법으로 영상을 구성하는 것이 가능하다. 또한 이러한 방법으로 인해 수어 사용자들의 동작 생략과 같은 경우에도 강인한 데이터 세트가 될 수 있다. 이러한 데이터 증강 과정을 거치면 총 1,151개의 비디오로 구성된 데이터 세트는 총 69,060개의 비디오로 구성된 데이터 세트로 증강된다.



<그림 3-11> 일정 간격 프레임 추출(위)

전체 프레임 무작위 추출(아래)의 예

제4장 실험 및 평가

제1절 실험 환경

본 장에서는 기존의 동작 인식 및 수어 관련 연구에 사용된 여러 기법을 실험해보고, 본 논문에서 제안하는 스켈레톤 영상 기반 방법과 비교해 본다. 또한 딥러닝 네트워크를 구성하는 여러 요소들에 대한 실험을 통해 한국 수어 인식에 적합한 데이터 및 파라메터들을 찾아내고 다른 한국 수어 인식 네트워크들과 비교하여 본 논문에서 제안하는 방법의 우수성을 보여준다.

본 논문에서 실험한 실험 환경은 <표 4-1>과 같다. 메모리가 32GB로 일반적인 PC사양보다는 높긴 하지만, 딥러닝 네트워크 서버로서는 매우 낮은 사양으로 볼 수 있다. 또한 무의미한 반복 훈련을 방지하고자 얼리스토핑을 사용하여, 훈련된 결과에 변화가 없을 경우 더 이상 훈련을 진행하지 않는다. 마지막으로 실험에 사용한 데이터 세트는 전체 데이터 세트를 훈련, 검증 및 시험 데이터로 분할하였으며, 그 비율은 8:1:1 이다.

<표 4-1> 실험에 사용된 PC 사양 및 환경 설정

부품 명	사양
CPU	Intel(R) Core(TM) I7-7700 @ 3.6GHz
GPU	NVIDIA GeForce GTX 1080(6 GB) x 2
Memory	32 GB
OS	Windows 10 Education x64
딥러닝 프레임워크	케라스 (텐서플로)
Earl Stopping	사용
데이터 세트 비율	8:1:1 (훈련 : 검증 : 시험)

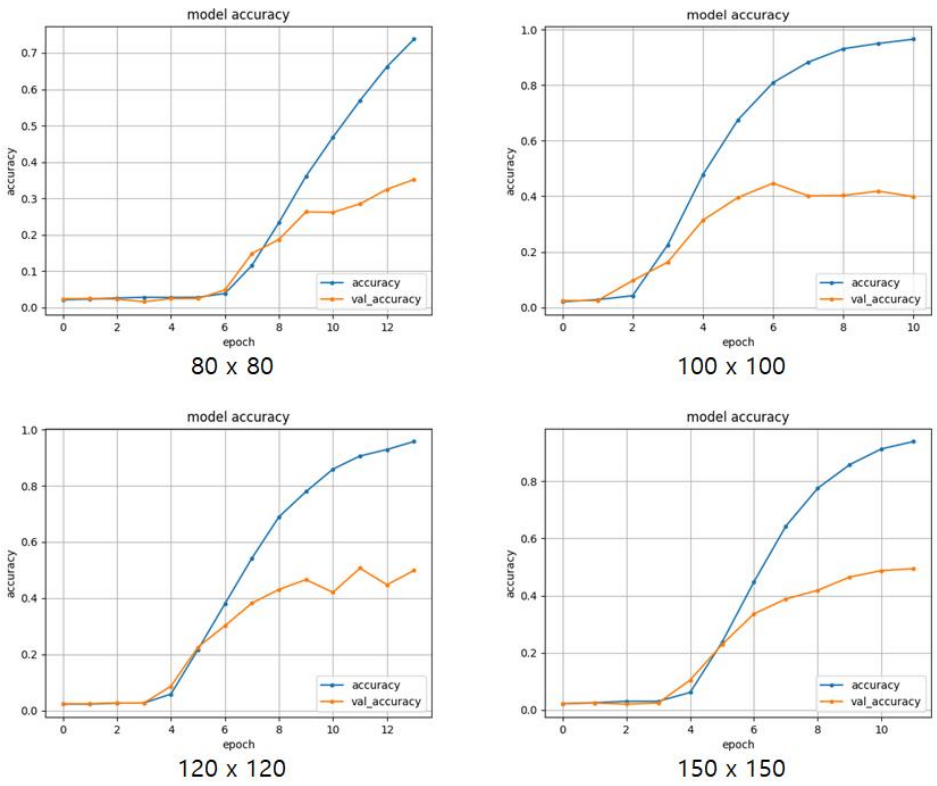
제2절 입력 데이터 구성 요소 실험 및 평가

본 절에서는 딥러닝 네트워크에 사용되는 입력 데이터를 구성하는 몇 가지 요소들에 대해 실험하고 평가한다. 동영상을 입력 데이터로 사용하게 되면, 정지 영상을 사용할 때와는 다르게 입력 영상에 몇 프레임을 사용할 것인지에 대한 문제가 발생한다. 정지 영상의 경우 단순히 이미지를 크게 하거나 줄이거나 하지만, 동영상의 경우는 다르다. 이미지를 크기와 입력 프레임의 수 사이에 선택의 문제가 발생한다. 가장 좋은 방법은 두가지 방법 모두를 수행하는 것이지만, 3D CNN의 경우 하드웨어 자원을 많이 소모하기 때문에 주어진 환경에서 최선의 선택이 필요한 문제이기도 하다.

본 절에서는 입력 이미지의 크기와 입력 데이터에 사용된 시퀀스의 프레임 개수를 조절하여 어떤 상황에서 더 좋은 결과가 나오는지 살펴본다.

1. 입력 영상 크기에 따른 실험 및 평가

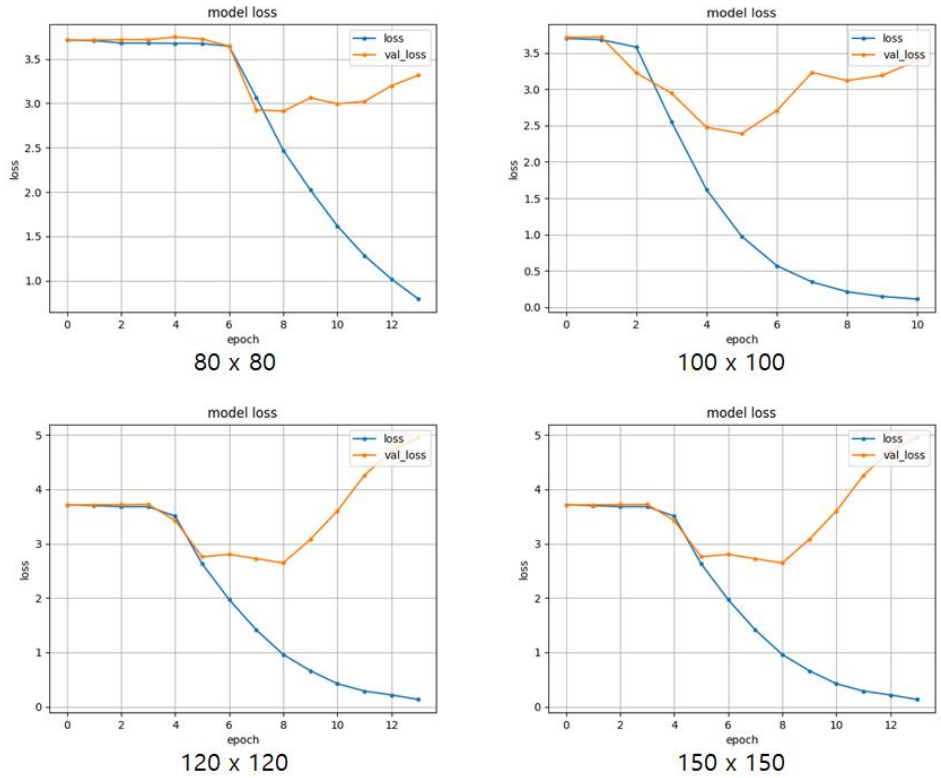
본 논문에서는 제안된 네트워크에 적합한 입력 영상의 크기를 찾기 위해 다양한 크기의 입력 데이터로 훈련을 했다. 사용된 데이터 세트는 KU 데이터 세트에서 단어당 15개의 영상을 가진 데이터 세트로 실험을 진행했다. 실험에 사용된 입력 영상의 크기는 80 x 80, 100 x 100, 120 x 120, 150 x 150이며, 모두 동일한 위치에서 10프레임을 추출하여 시퀀스를 만들었다. 스켈레톤 영상이 아닌 RGB 영상을 사용했으며, 무작위 프레임추출을 제외한 이미지 증강 작업은 수행 되었다. 실험에는 C3D[36] 네트워크를 사용했다. C3D 네트워크는 8개의 3D 컨볼루션 레이어와 5개, 최대 풀링 레이어, 2개의 완전 연결 레이어로 구성된다. Adam 옵티마이저를 사용했으며, 학습률은 0.00001, 감소율은 0.0001로 설정했다.



<그림 4-1> 입력영상 크기에 따른 훈련 결과(정확도)

<그림 4-1>은 각 영상 크기의 훈련에 대한 정확도를 그래프로 보여준다. 각 그래프의 Y축은 정확도를 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 정확도, 적색은 검증 데이터에 대한 정확도를 나타낸다. 그래프 상으로는 훈련 시 120 x 120의 입력 영상 크기가 가장 좋은 결과를 보여주고 80 x 80의 입력 영상 크기가 가장 나쁜 결과를 보여준다. 150 x 150의 입력 영상크기의 경우 100 x 100의 입력 영상 크기와 비슷한 결과를 보여준다. <그림 4-2>는 손실값을 보여준다. 각 그래프의 Y축은 손실값을 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 손실, 적색은 검증 데이터에 대한 손실값을 나타낸다. 120 x 120의 입력 영상

크기의 결과가 가장 안 좋으며, 100 x 100의 입력 영상 크기의 결과가 가장 좋은 것을 확인할 수 있다.



<그림 4-2> 입력영상 크기에 따른 훈련 결과(손실값)

<표 4-2>는 훈련된 네트워크에 시험 데이터를 사용하여 실험한 결과 정확도와 손실값을 표로 보여준다. 시험 데이터는 훈련 데이터 및 검증 데이터와 중복이 되지 않는 영상을 사용했다. 그 결과 예측한 대로 정확도는 150 x 150의 입력 영상 크기의 결과가 가장 좋았으며, 예상과는 다르게 손실값은 100 x 100의 결과가 가장 좋게 나왔다. 80 x 80의 입력 영상 크기는 정확도에서 가장 안 좋은 결과를 보여준다.

전반적으로 예상한 대로 입력 영상의 크기가 클수록 높은 정확도를 가지 가지는 것을 알 수 있다. 하지만 손실값의 크기가 모든 결과에서 5에 가깝거나 더 높게 나온 것을 보았을 때 해당 실험은 큰 의미를 가지지는 못한다고 볼 수 있다.

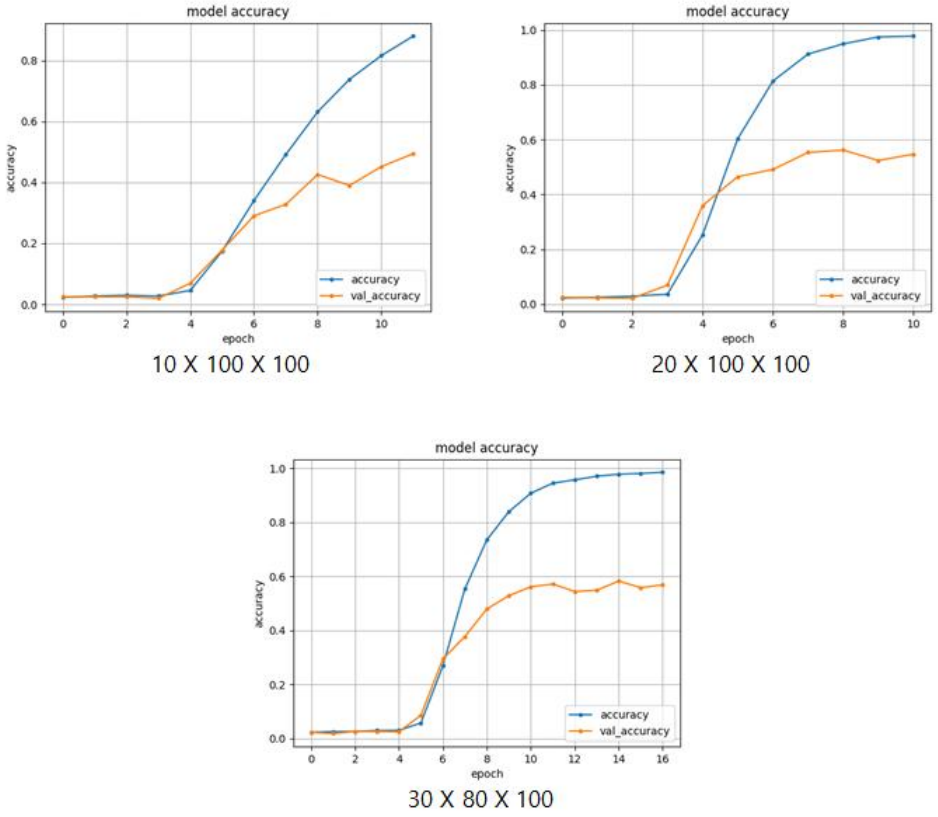
<표 4-2> 시험 데이터를 사용한 입력 영상 크기 별 실험 결과

입력 영상크기	정확도	손실값
80 x 80	0.41	5.26
100 x 100	0.45	**4.62**
120 x 120	0.49	5.95
150 x 150	**0.51**	4.93

2. 입력 시퀀스의 길이에 따른 실험 및 평가

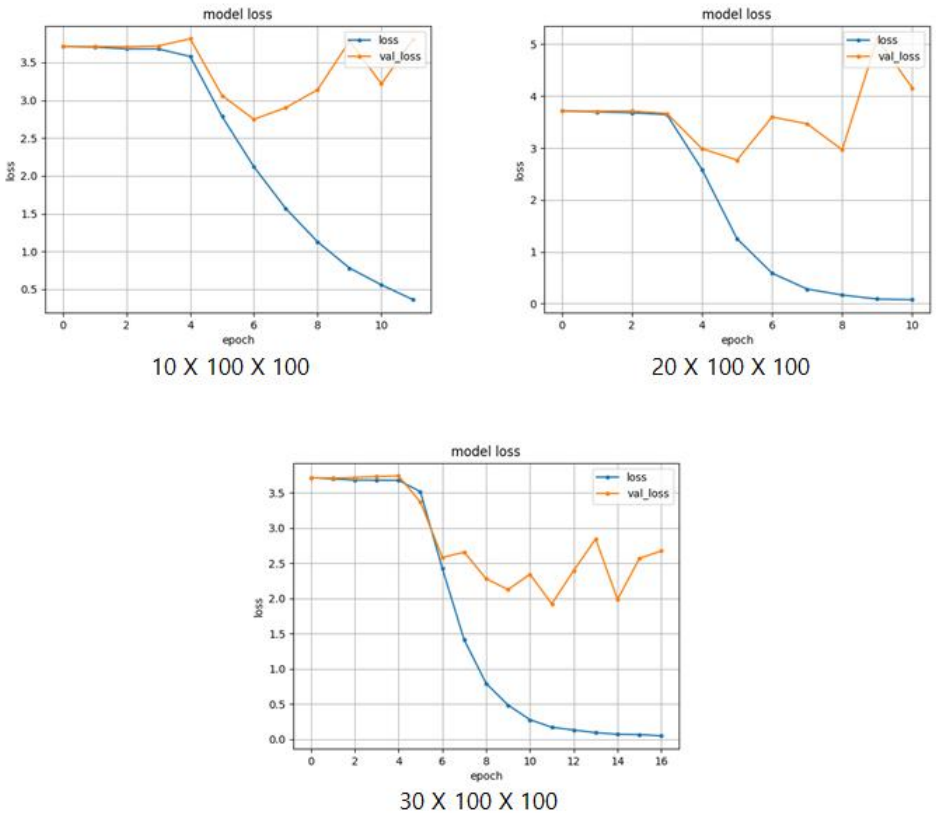
입력 시퀀스의 길이는 정확도에 큰 영향을 끼친다. 네트워크에 보다 많은 프레임을 입력 시퀀스로 사용하게 되면 결과는 좋아 질 수밖에 없다. 하지만 실제 딥러닝 네트워크를 훈련시킬 때 하드웨어의 자원은 무한하지 않다. 때문에 제한된 하드웨어 환경에서 적절한 입력 시퀀스의 길이를 선정하는 것은 매우 중요하다. 하지만 입력 시퀀스의 길이만을 가지고 실험을 하게 되면, 네트워크의 훈련 결과는 입력 시퀀스의 길이에 비례하기 때문에 무의미한 데이터가 된다. 따라서 본 논문에서는 시퀀스의 길이와 이미지의 크기를 복합으로 사용하여 훈련에 적절한 시퀀스 길이를 구하고자 한다. 따라서 첫번째 실험에서는 앞선 입력영상의 크기에 관한 실험 결과 100 x 100크기의 입력영상을 대상으로 시퀀스 길이를 10, 20, 30프레임으로 만들어 시퀀스 길이에 따른 결과를 확인하였다.

<그림 4-3>는 각 시퀀스의 길이(10장, 20장, 30장)에 대한 훈련에 대한 정확도 결과를 보여준다. 각 그래프의 Y축은 정확도를 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 정확도, 적색은 검증 데이터에 대한 정확도를 나타낸다. 훈련 결과는 시퀀스 길이가 30프레임의 경우 가장 좋은 결과를 보여주며, 예상대로 프레임이 늘어날수록 점점더 좋은 결과를 보여준다. <그림 4-4>는 각 시퀀스 길이에 대한 훈련의 손실값 결과를 보여준다. 각 그래프의 Y축은 손실값을 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 손실값, 적색은 검증 데이터에 대한 손실값을 나타낸다. 시퀀스 길이가 30프레임일 경우 가장 좋은 결과를 보여준다.



<그림 4-3> 100 x 100 영상 시퀀스 길이에 따른 훈련 결과(정확도)

<표 4-3>에서 보여주는 훈련에 사용하지 않은 시험데이터로 확인한 정확도와 손실값을 확인해 보면, 입력 영상의 시퀀스 길이가 가장 긴 30 프레임일 경우 가장 높은 정확도와 가장 낮은 손실값의 결과를 가지는 것을 볼 수 있다.



<그림 4-4> 100 x 100 영상 시퀀스 길이에 따른 훈련 결과(손실값)

<표 4-3> 100 x 100 크기의 시험 데이터를 사용한

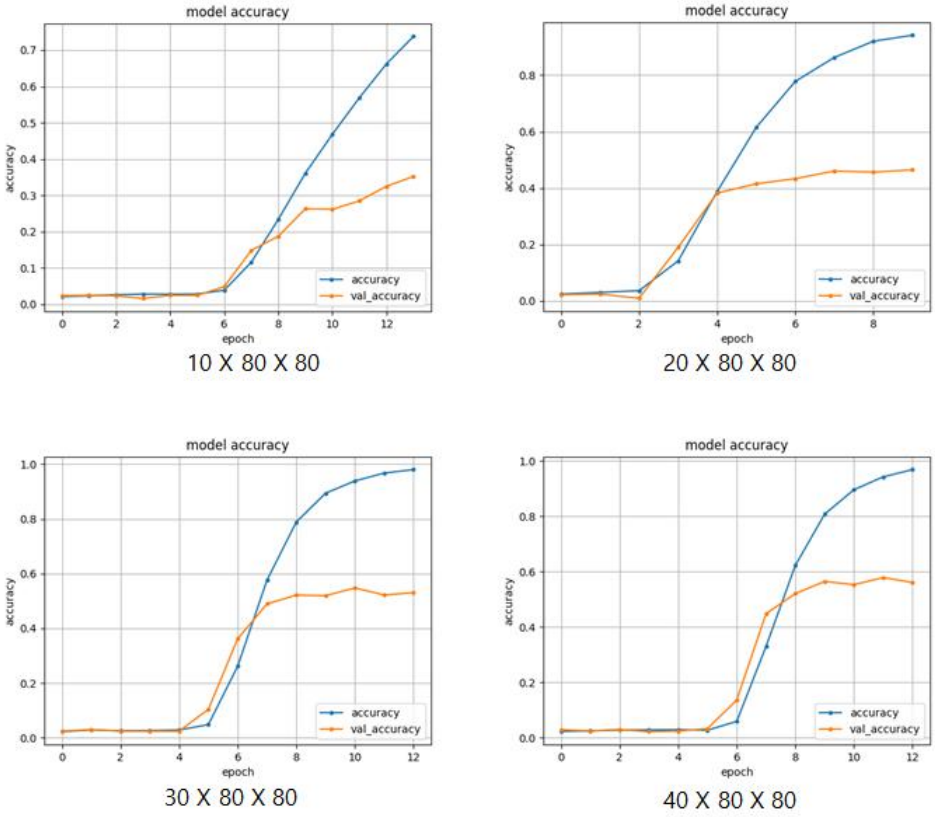
시퀀스 길이에 따른 정확도

입력 영상크기	정확도	손실값
10 x 100 x 100	0.47	3.56
20 x 100 x 100	0.52	5.73
30 x 100 x 100	**0.58**	**2.74**

100x100 크기의 영상 실험의 결과 가장 긴 시퀀스에 대하여 가장 좋은 결과가 나왔기 때문에 더 긴 시퀀스에 대하여 성능을 확인할 필요가 있으나 실험환경의 제약으로 40장에 대한 실험을 수행하는 것이 불가능하였다. 이에 대한 대안으로 80 x 80크기의 입력영상을 시퀀스 길이를 10, 20, 30, 40 프레임으로 만들어 시퀀스 길이에 따른 결과를 다시 확인하였다.

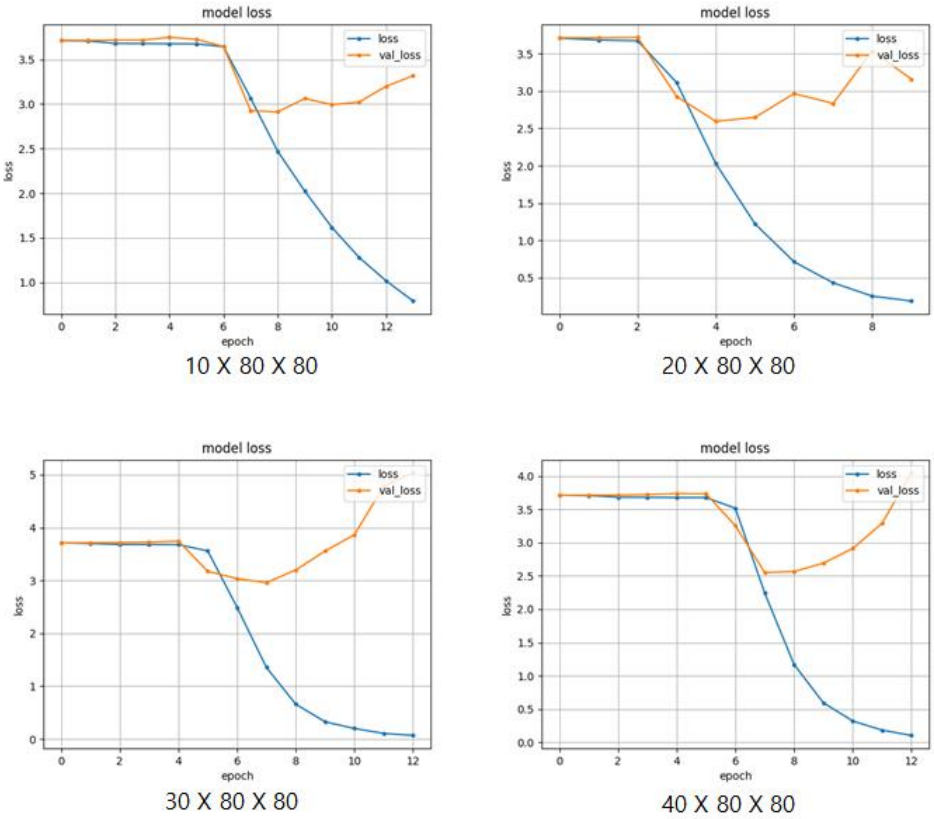
<그림 4-5>는 각 시퀀스의 길이(10장, 20장, 30장, 40장)에 대한 훈련에 대한 정확도 결과를 보여준다. 각 그래프의 Y축은 정확도를 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 정확도, 적색은 검증 데이터에 대한 정확도를 나타낸다. 훈련 결과는 시퀀스 길이가 30프레임 및 40프레임일 경우 유사한 결과를 보여주며, 예상대로 10프레임과 20프레임은 안 좋은 결과를 보여준다.

<그림 4-6>는 각 시퀀스 길이에 대한 훈련의 손실값 결과를 보여준다. 각 그래프의 Y축은 손실값을 나타내고 X축은 에포크를 의미하며, 청색은 훈련 데이터에 대한 손실값, 적색은 검증 데이터에 대한 손실값을 나타낸다. 시퀀스 길이가 30프레임일 경우와 40프레임일 경우 좋지 않은 결과를 보여주며, 시퀀스 길이가 20프레임일 경우 가장 좋은 결과를 보여준다.



<그림 4-5> 80 x 80 영상 시퀀스 길이에 따른 훈련 결과(정확도)

하지만 <표 4-4>에서 보여주는 훈련에 사용하지 않은 시험데이터로 확인한 정확도와 손실값을 확인해 보면, 입력 영상의 시퀀스 길이가 가장 긴 40 프레임일 경우 가장 높은 정확도와 가장 낮은 손실값의 결과를 가지는 것을 볼 수 있다. 따라서 예상한 바와 같이 시퀀스의 길이가 더 길어진다면 네트워크의 성능이 향상되는 것을 확인할 수 있었다.



<그림 4-6> 80 x 80 영상 시퀀스 길이에 따른 훈련 결과(손실값)

<표 4-4> 80 x 80 크기의 시험 데이터를 사용한

시퀀스 길이에 따른 정확도

입력 영상크기	정확도	손실값
10 x 80 x 80	0.41	5.26
20 x 80 x 80	0.49	5.36
30 x 80 x 80	0.52	5.32
40 x 80 x 80	**0.59**	**4.30**

3. 입력 영상 크기 및 시퀀스 길이에 따른 네트워크 파라미터 비교

앞 절에 주어진 실험에서 입력 영상의 크기와 시퀀스 길에 따른 실험 결과를 살펴보았다. 하지만 실제 네트워크를 구성하고 실험을 하는 경우, 하드웨어의 자원으로 인해 여러가지 제약이 발생하게 된다. 딥 러닝 네트워크의 경우 입력 영상이 크면 클수록, 입력 영상의 길이가 길면 길수록 네트워크에서는 더 많은 파라미터들을 사용하게 되고, 이는 더 많은 하드웨어 자원을 필요로 하게 된다.

<표 4-5> 각 실험에 따른 사용 파라미터 비교

	10 프레임	20 프레임	30 프레임	40 프레임
80 x 80	82,881,321	135,310,121	187,738,921	240,167,721
100 x 100	105,949,993	181,447,465	256,944,937	332,442,409
120 x 120	133,212,969	235,973,417	338,733,865	441,494,313
150 x 150	200,321,833	370,191,145	540,060,457	709,929,769

<표 4-5>는 본 연구에서 실험한 입력 영상의 크기와 입력 시퀀스 길에 따른 네트워크 파라미터 사용량을 나타낸다. 150 x 150 크기에 10프레임의 길이를 가지는 시퀀스를 사용하는 네트워크는 약 2억개의 파라미터가 필요하다. 동일한 크기에서 프레임을 10프레임씩 늘려 갈때마다 약 1억7천개씩의 파라미터 (약84%)가 추가로 필요하게 된다. 하지만 입력 영상의 크기가 80 x 80 크기의 경우에는 약 5천만개의 파라미터(64%)가 추가로 필요하게 된다. 이로 인해 좀더 작은 크기의 영상에서 프레임의 길이를 늘리는 것이 더 적은 파라미터를 사용하는 것을 알 수 있다. 본 논문에서는 실험환경의

제약으로 인하여, <표 4-5>에서 녹색 부분으로 색칠되어 있는 블록들, 즉 3억개 이하의 파라미터 수를 갖는 구성에 대하여만 실험이 가능하였다.

앞선 실험의 결과 입력 영상의 크기는 100 x 100이 가장 좋았으며, 시퀀스의 길이는 40프레임이 가장 좋은 결과를 보였다. 이러한 결과대로라면 입력영상의 크기는 100 x 100을 가지는 40 프레임 길이의 시퀀스를 입력 영상으로 사용하는 것이 가장 좋은 실험 결과를 보여줄 것으로 예상되지만 100 x 100크기에 40프레임 길이의 시퀀스는 주어진 환경에서 지원할 수 없는 하드웨어 자원을 요구한다. 때문에 본 논문에서는 네트워크에서 사용하기 위한 최종 입력 시퀀스로 크기는 100 x 100으로 유지하고, 시퀀스의 길이는 30프레임을 가지는 입력 시퀀스를 사용하기로 한다.

제3절 수어 데이터 세트를 사용한 실험 및 평가

본 절에서는 본 논문에서 제안한 KU수어 데이터 세트에 대한 실험 및 평가를 기술한다. 우선 데이터 세트의 크기가 정확도에 미치는 영향을 파악하고, 다른 한국 수어 데이터 세트인 KETI 데이터 세트를 본 논문에서 제안하는 방법과 KETI 데이터 세트에서 제안하는 방법을 비교해 본다. 본 절의 실험에서는 기본적으로 KU 데이터 세트를 소규모로 구성한 KU 데이터 세트 15와 C3D 네트워크를 사용한다. 사용된 데이터 세트는 기본적으로 100 x 100 크기와 30프레임 길이의 입력 영상 시퀀스를 사용하나, 상황에 따라 다른 크기 및 길이를 가진 입력 영상 시퀀스를 사용한다. KU 데이터 세트 15는 41개 단어의 동영상으로 구성되어 있으며, 각 단어 당 영상의 수는 15개로 총 영상 615로 구성된 작은 데이터 세트다.

1. 데이터 증강 및 변환에 따른 실험 및 평가

데이터 세트의 크기는 딥러닝 네트워크를 훈련하는데 매우 중요한 요소다. 데이터 세트가 부족하면 딥러닝 네트워크가 정상적으로 훈련이 되지 않는다. 본 논문에서는 소규모로 구성한 KU 수어 데이터 세트 15를 사용하여, 본 논문에서 제안하는 데이터 증강 방법에 대한 성능을 보여준다. 또한 RGB 영상이 아닌 스켈레톤 영상을 입력 영상으로 사용했을 때의 성능 향상을 보여준다. KU 데이터 세트 15는 데이터 증강 기법을 통해 36,900개의 영상으로 증강된다.

본 실험은 2가지에 대한 실험을 진행한다. 첫째, 입력 영상의 유형에 따른 결과를 확인해 본다. 둘째, 영역 분할 후 분할 영역내 프레임 추출방법과 전체 프레임 내 랜덤 추출 방법을 비교하고 결과를 살펴본다.

첫번째 실험인 입력 영상의 유형에 따른 실험은 RGB 영상과 스켈레톤 영상을 사용한다. 100 x 100크기의 영상을 사용하며, 10, 20, 30 프레임의 길이를 갖는 입력 영상 시퀀스를 사용하여 실험을 진행했다. RGB 영상의 경우 복잡한 배경, 다양한 사람 및 의상으로 구성된 영상들이다. RGB 영상의 경우, 공간적 특징을 추출하고자 할 때, 배경으로 인한 잘못된 특징을 추출할 수 있다. 또한 촬영자의 의상 및 동작에 따라 동작이 배경으로 인하여 인식되지 않기도, 전혀 엉뚱한 특징이 추출되기도 한다. 때문에 복잡한 배경, 인물, 의상을 포함하는 데이터의 경우 높은 정확도를 기대하기는 어려운 것 또한 사실이다. 본 논문에서 한 실험의 결과도 마찬가지로 높지 않은 정확도를 보여준다. <표 4-6>에서 보는것과 같이 입력 영상으로 RGB 영상을 사용한 경우, 100 x 100 크기의 입력 영상을 사용했을 때 시퀀스 길이에 따라 약 47~58%의 정확도를 보여주며, 손실값 또한 5에 가까운 값을 보여

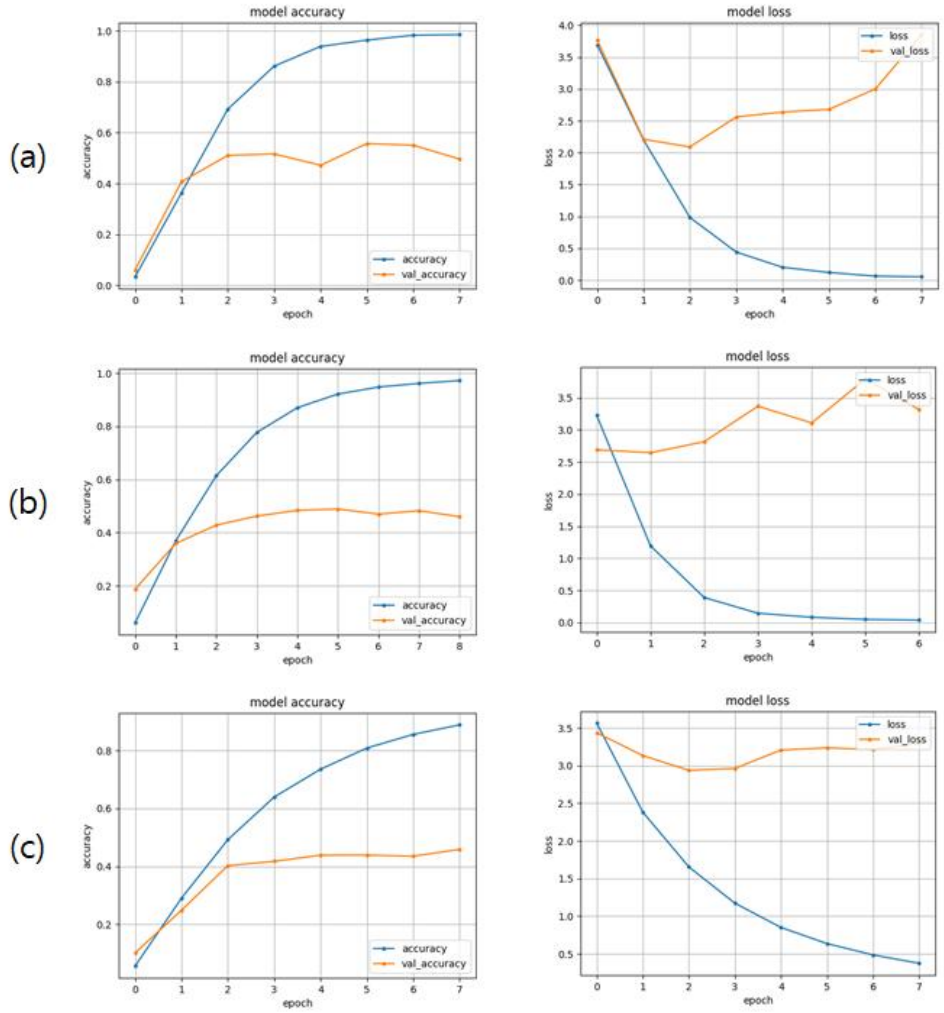
준다. 스켈레톤 영상의 경우 위에 언급한 문제점들이 해결된 영상이다. 영상 내 사람의 스켈레톤 영상만 존재하기 때문에 복잡한 배경이 없으며, 같은 이유로 인한 복잡한 의상 또한 존재 하지 않는다. 또한 다양한 사람들로 구성된 데이터라 해도 스켈레톤 영상으로 변환 시 대부분이 비슷한 형태를 가지는 영상으로 변하게 된다. 배경이 존재하지 않기 때문에 사람의 움직임을 파악하기 쉬우며 이는 결과적으로 정확도의 상승이라는 결과가 만들어지게 된다. <표 4-6>에서 보는것과 같이 스켈레톤 영상을 사용한 경우 RGB영상을 사용한 것에 비해 정확도는 약 2~5% 정도 상승한 것을 볼 수 있으며, 손실값 또한 2.5~0.5 정도가 줄어든 것을 볼 수 있다. 이런 이유로 인해 스켈레톤 영상이 RGB 영상보다 더 좋은 결과를 보여준다는 것을 알 수 있다.

<표 4-6> 시험 데이터를 사용한 영상 비교 실험 결과

시퀀스 길이	RGB		스켈레톤	
	정확도	손실값	정확도	손실값
10 프레임	0.47	3.56	0.52	3.15
20 프레임	0.52	5.73	0.57	3.20
30 프레임	0.58	4.80	0.60	3.32

앞에서 확인한 결과에 따라, 랜덤 프레임 추출 방법에 대한 성능 실험은 스켈레톤 영상으로 진행된다. 랜덤 프레임 추출 실험은 3가지 데이터를 가지고 비교하게 된다. 프레임 추출을 사용하지 않았을 때의 결과, 프레임을 특정 개수로 분할하고 분할한 영역에서 무작위로 프레임을 추출하는 방법 및 전체 프레임내에서 무작위로 프레임을 추출하는 방법의 결과를 비교하게

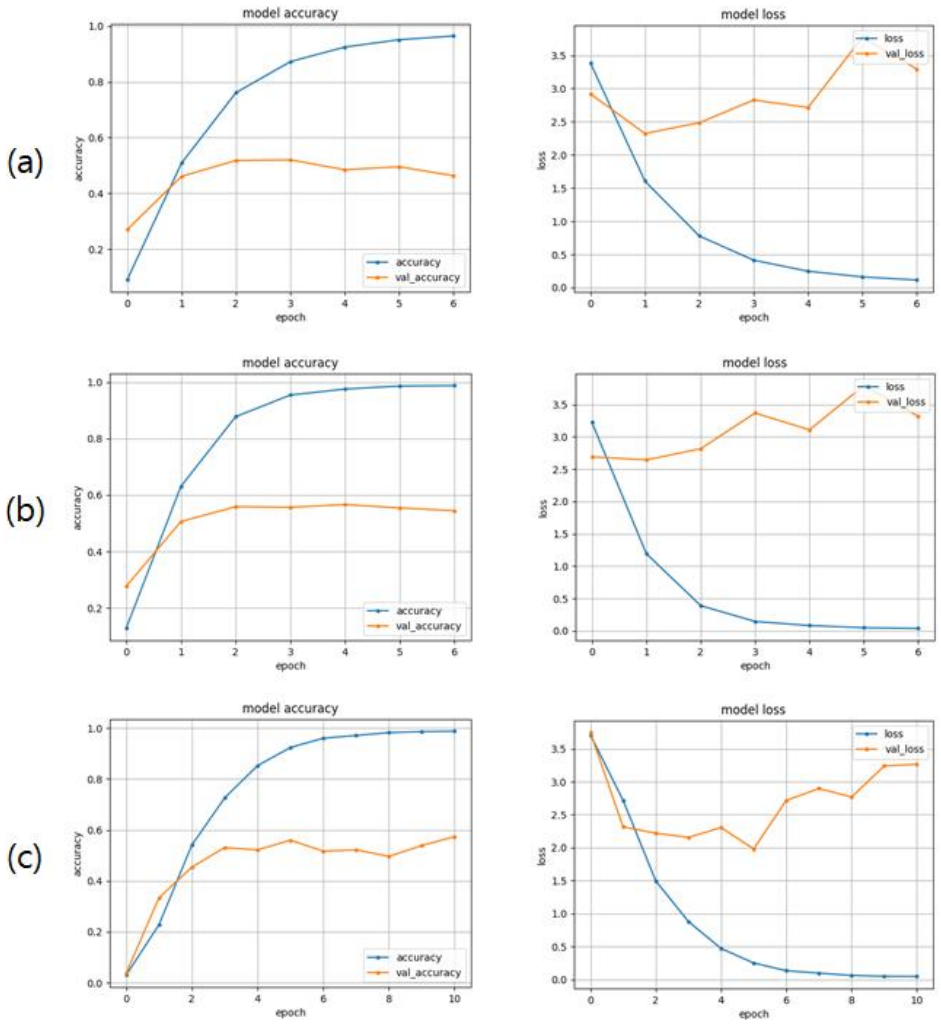
된다. <그림 4-7>는 입력 시퀀스의 길이가 10프레임일 때의 랜덤 프레임 추출방법에 대한 결과를 보여준다. 마찬가지로 <그림 4-8>과 <그림 4-9>는 각각 시퀀스길이가 20 및 30프레임일 때의 랜덤 프레임 추출 방법에 대한 정확도 와 손실값 결과를 보여준다.



<그림 4-7> 시퀀스 길이 10프레임의 랜덤 프레임 추출 실험 결과

(a) 랜덤 프레임 추출 사용 안함, (b) 분할 영역 내 무작위 추출,

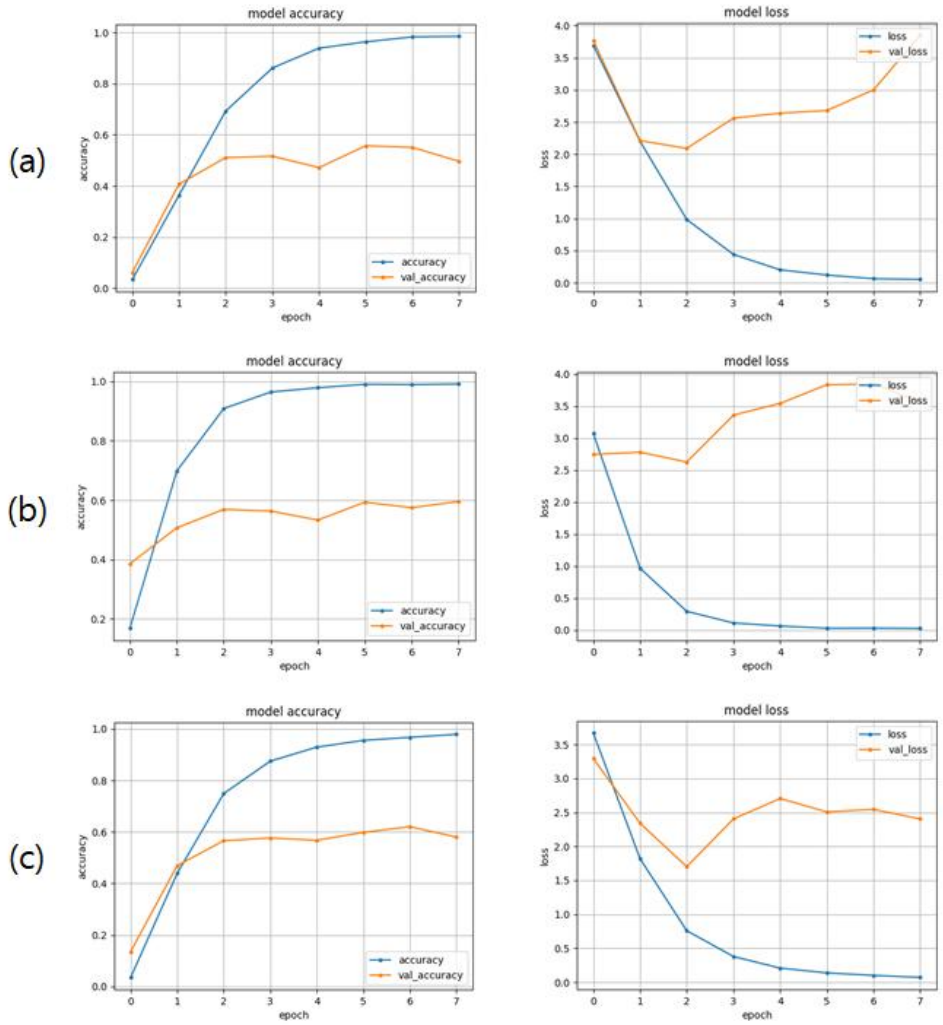
(c) 전체 프레임내 무작위 추출



<그림 4-8> 시퀀스 길이 20프레임의 랜덤 프레임 추출 실험 결과

(a) 랜덤 프레임 추출 사용 안함, (b) 분할 영역 내 무작위 추출,

(c) 전체 프레임내 무작위 추출



<그림 4-9> 시퀀스 길이 30프레임의 랜덤 프레임 추출 실험 결과

(a) 랜덤 프레임 추출 사용 안함, (b) 분할 영역 내 무작위 추출,

(c) 전체 프레임 내 무작위 추출

<그림 4-7>부터 <그림 4-9>까지의 그래프를 살펴보면 시퀀스의 길이가 길어 질수록 정확도가 좀더 빠르게 상승하는 것을 볼 수 있다. 또한 손실값의 그래프도 좀더 빠르게 내려가는 것을 확인할 수 있다. <표 4-7>은

위와 같이 훈련된 네트워크에 시험 데이터 세트를 이용해 확인한 결과를 나타낸다. 그 결과 전체 프레임 내 무작위 추출하는 방법이 분할 영역 내 무작위 추출하는 방법보다 더 높은 정확도를 보여주는 것을 확인할 수 있었다. 시퀀스의 길이가 10, 20프레임일 경우, 분할 영역 내 무작위 추출과 전체 프레임내 무작위 추출 방법은 약 2~3% 정도의 정확도 차이가 발생한다. 하지만 시퀀스 길이 30프레임의 경우 약 9%의 정확도 차이가 발생한다. 이는 시퀀스의 길이가 길어질수록 전체 프레임 내 무작위 추출 방법의 성능이 더 좋아진다는 걸 보여준다. 손실값 또한 랜덤 프레임 추출을 사용하지 않은 경우보다 약 25%의 성능 향상을 보여준다.

<표 4-7> 시험 데이터를 사용한 랜덤 프레임 추출 실험 결과

시퀀스 길이	무작위 추출 사용 안함		분할 영역 내 무작위 추출		전체 프레임 내 무작위 추출	
	정확도	손실값	정확도	손실값	정확도	손실값
10	0.52	3.15	0.59	2.88	0.61	2.50
20	0.57	3.20	0.61	2.75	0.64	2.39
30	0.60	3.32	0.64	2.50	**0.73**	**2.07**

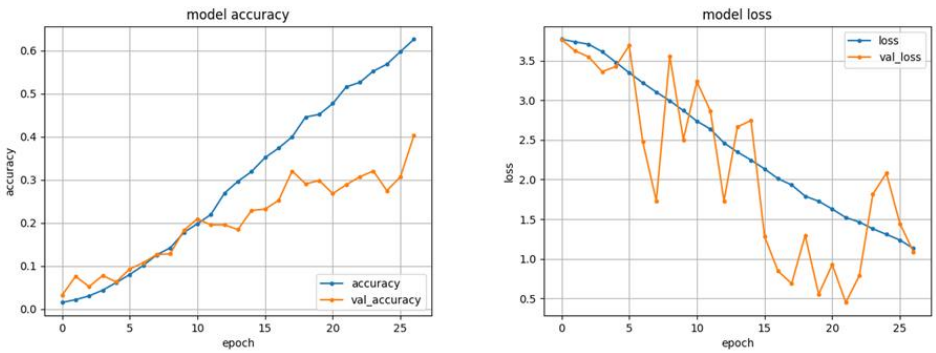
2. 네트워크 구성에 따른 실험 및 평가

본 실험에서는 본 논문에서 제안한 KU 수어 단어 인식 네트워크 모델을 비롯해 C3D[36], LRCN등 다양한 모델을 이용한 실험을 살펴본다. 사용한 데이터 세트는 KU 데이터세트 15를 사용하며, 입력 영상의 크기는 100 x 100, 시퀀스 길이는 30 프레임이다. 본 실험에서 사용한 네트워크는 다음

과 같다.

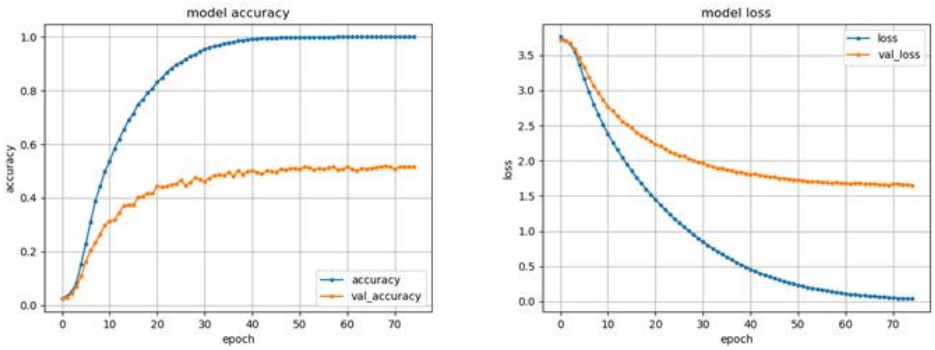
- LRCN
- Merge 3D CNN
- C3D
- Compact C3D
- Simple 3D CNN
- Middle 3D CNN
- KU 수어 단어 인식 네트워크

LRCN 모델은 입력 시퀀스의 각 프레임을 ImageNet 가중치를 사용한 InceptionV3 모델을 사용해 특징 값을 추출한다. 추출된 특징 값들은 LSTM 레이어의 입력 값으로 사용되며 완전 연결 레이어로 전달된다. LRCN 실험 결과는 <그림 4-10>과 같다. 그래프에서 보이는것과 같이 정확도, 손실값 모두 매우 좋지 않은 결과를 보여준다. 특히 손실값의 경우 지그재그 형태로 매우 요동치는 모습을 볼 수 있을 정도로 매우 안 좋은 결과를 보여준다.



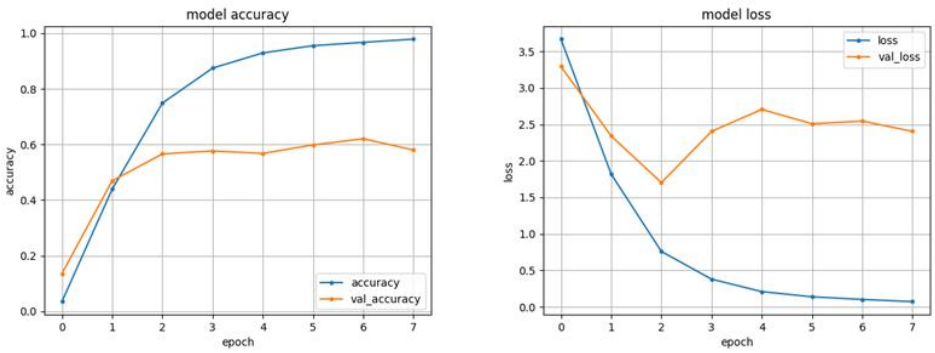
<그림 4-10> LCRN 실험 결과

Merge 3D CNN은 하나의 영상으로부터 서로 다른 시퀀스 2개를 추출하여 입력 영상으로 사용한다. 하나의 영상은 일정 간격으로 프레임을 추출한 영상이고, 또 다른 영상은 랜덤 프레임 추출로 만들어진 영상을 사용한다. 각 입력 영상은 3D 컨볼루션 레이어와 3D 풀링 레이어를 3번 지나 병합되고, 병합된 특징들은 완전 연결 레이어로 전달된다. <그림 4-11>는 Merge 3D CNN의 실험 결과를 보여준다. LRCN과는 다르게 안정적인 그래프를 보여준다.



<그림 4-11> Merge 3D CNN 실험 결과

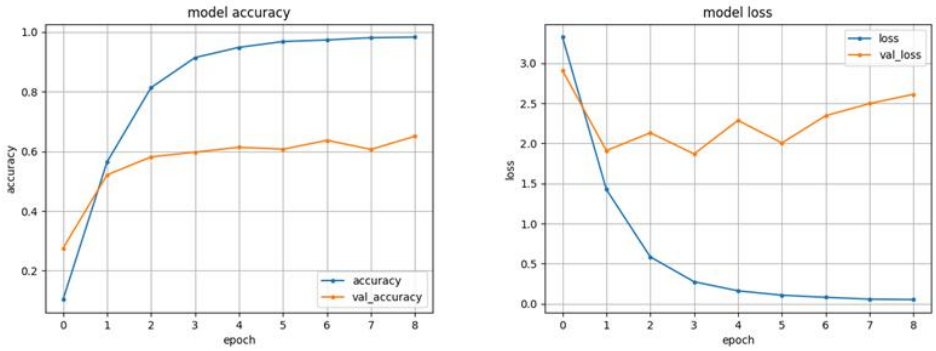
<그림 4-12>에서 사용된 C3D 모델은 동작 인식 연구에서 많이 사용되는 모델이며, 높은 정확도를 보여주는 모델이다. C3D의 일반적인 입력 영상으로 128 x 171 크기, 16프레임의 영상을 사용한다. 하지만 본 논문의 실험에서는 100 x 100 크기, 30프레임의 영상을 입력으로 사용한다. C3D 모델은 컨볼루션 레이어와 풀링 레이어로 구성된 5개의 그룹과 하나의 제로 페딩 레이어, 2개의 완전 연결 레이어로 구성되는 모델이다. C3D의 경우 <그림 4-10>에서 보는것과 같이 정확도와 손실값 모두에서 Merge 3D CNN보다 좋은 결과를 보여준다.



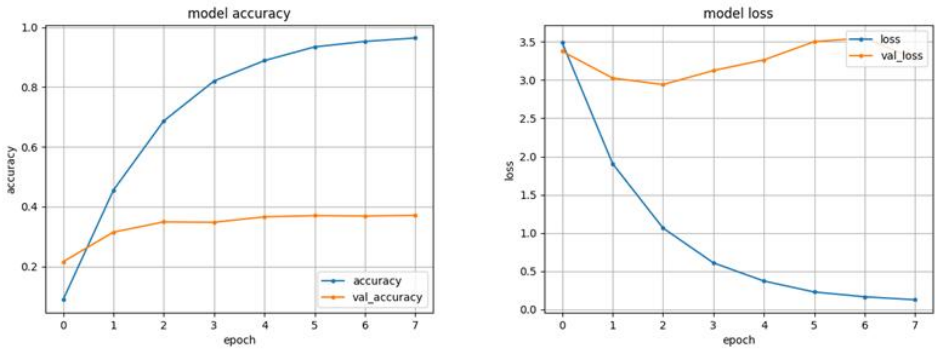
<그림 4-12> C3D 실험 결과

<그림 4-13>은 Compact C3D 모델의 결과를 보여준다. 이는 기존 C3D 모델에서 하나의 컨볼루션 레이어와 하나의 최대 풀링 레이어로 구성된 마지막 그룹을 제외하고 제로 패딩 레이어를 제외한 모델이다. 또한 완전 연결 레이어의 구성을 4096개에서 1024개로 변경하여 3D 컨볼루션 영역에서 완전 연결 영역으로 넘어가는 특징 및 가중치에 대한 개수를 증가시켰다. Compact C3D 모듈은 정확도 측면에서는 C3D보다 좋은 결과를 보여주며, 손실값 또한 C3D와 비슷한 결과를 보여준다. 이로 인해 완전 연결 영역에 전달되는 파라미터의 개수가 많은 것은 정확도 측면에서 오히려 결과가 좋게 나온다는 것을 알 수 있다.

<그림 4-14>에서 보여주는 Simple 3D CNN 모델은 1개의 컨볼루션 레이어와 1개의 최대 풀링 레이어로 구성된 그룹 3개와 2개의 컨볼루션 레이어와 1개의 최대 풀링 레이어로 구성된 하나의 그룹, 그리고 3개의 완전연결 레이어로 구성되어 있다. Simple 3D CNN은 다른 모델에 비해 적은 필터의 개수와 낮은 깊이를 지닌 모델이다. 실험의 결과 또한 낮은 깊이로 인해 <그림 4-14>에서 보여주는 바와 같이 낮은 정확도를 보여준다.



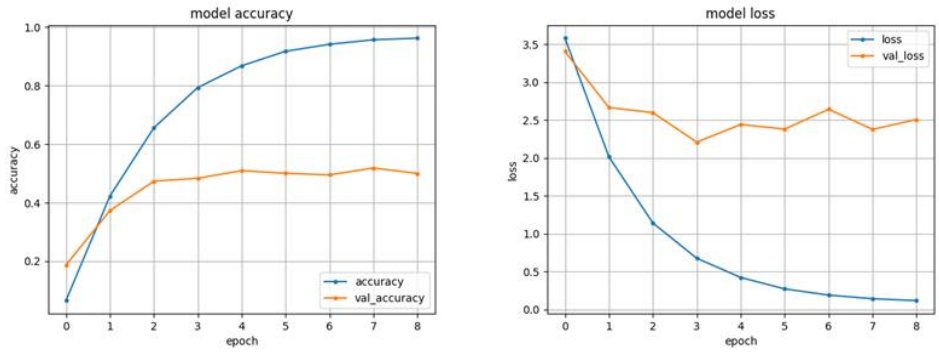
<그림 4-13> Compact C3D 실험 결과



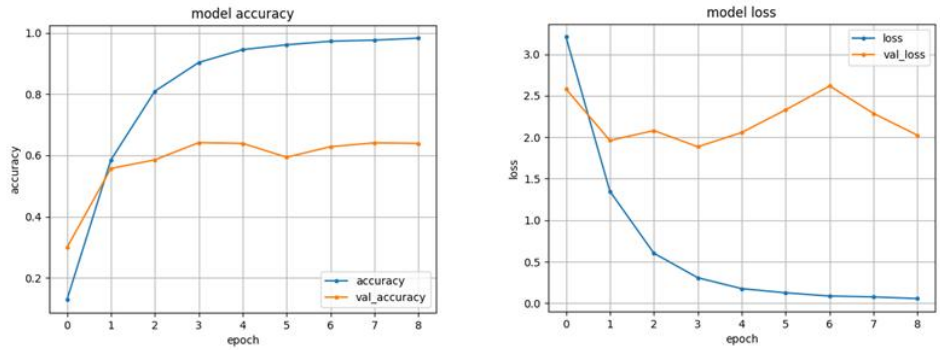
<그림 4-14> Simple 3D CNN 실험 결과

<그림 4-15>은 Middle 3D CNN 모델의 실험 결과를 보여준다. Middle 3D CNN 모델은 1개의 컨볼루션 레이어와 1개의 최대 풀링 레이어로 구성된 4개의 그룹과 2개의 완전 연결 레이어로 구성되어 있다. 각 컨볼루션 레이어의 필터 개수는 Simple 3D CNN보다 2배 더 많은 레이어다. 대신 완전 연결 레이어를 하나 감소시켜 전체적으로 레이어의 깊이는 Simple 레이어와 동일한 깊이를 가진다. Middle 3D CNN모델의 실험 결과는 <그림 4-15>과 같다. 동일한 깊이의 모델인 Simple 3D CNN 모델보다 높은 정확도를 확인할 수 있다. 이 실험으로 인해 완전 연결 레이어 보다는 컨볼루션

레이어가 정확도에 더 큰 영향일 미치는 것을 알 수 있다.



<그림 4-15> Middle 3D CNN 실험 결과



<그림 4-16> KU 수어 단어 인식 네트워크 실험 결과

마지막으로 KU 수어 단어 인식 네트워크 모델은 본 논문에서 제안하는 모델이며, C3D[36]의 구조를 참조하여 확장한 모델이다. KU 수어 단어 인식 네트워크 모델은 100 x 100의 해상도와 30 프레임의 길이를 가지는 입력영상에 적합하도록 네트워크를 수정 및 최적화했다. 또한 완전 연결 레이어로 진행되기 전에 제로 패딩 레이어는 삭제하고 2D 컨볼루션 레이어를 추가하여 정확도를 높이고자 했다. <그림 4-16>는 KU 수어 단어 인식 네트워크 모델의 실험 결과를 보여준다. 그래프 상으로 보이는 정확도와 손실

값은 C3D와 비슷한 값을 보여준다.

<표 4-8>은 지금까지 실험 네트워크 모델에 시험 데이터를 적용하여 실험한 결과를 보여준다. KU 수어 단어 인식 네트워크의 정확도가 74%로 가장 높은 정확도를 보여주며, LRCN이 49%로 가장 낮은 정확도를 보여준다.

<표 4-8> 시험 데이터를 사용한 네트워크 모델 실험 결과

모델 명	정확도	손실값
LRCN	0.49	**0.64**
Merge 3D CNN	0.62	1.31
C3D	0.73	2.07
Compact C3D	0.74	1.69
Simple 3D CNN	0.56	2.23
Middle 3D CNN	0.65	1.68
KU 수어 단어 인식 네트워크	**0.76**	1.39

3. 데이터 세트의 크기에 따른 실험 및 평가

데이터 세트의 크기가 아무리 크다고 해도, 각각의 클래스를 구성하는 데이터의 수가 적으면 그 또한 정상적인 데이터 세트라 할 수 없다. 본 논문에서는 딥러닝 네트워크가 높은 정확도를 가지기 위해 필요한 클래스 당 데이터의 수를 구하기 위해 KU 수어 데이터 세트로 진행한 실험을 보여준다. 앞선 실험에서 데이터 증강의 효과를 확인했기 때문에 본실험은 모두 스켈레톤 영상을 사용하며, 데이터 증강기법 또한 모두 적용한 상태로 측정했다. 또한 앞선 실험에서 결과가 가장 좋은 KU 수어 단어 인식 네트워크

를 사용하며, KU 수어 단어 인식 네트워크에 추가된 2D 컨볼루션 레이어의 효과를 살펴본다.

본 실험에서 사용한 데이터 세트들은 모두 단어 당 영상의 수를 조절하여 만들었다. 데이터 세트들은 각각 10, 15, 28개의 단어 당 영상의 수를 가지고 있으며, 이는 <표 4-9>과 같은 구성을 가진다. 각 데이터 세트는 KU 데이터 세트 10을 기준으로 하며, 그 외 데이터 세트들은 기준 데이터 세트에 포함되지 않은 새로운 영상 세트(41개의 단어 영상)를 추가했다. 이는 차후에 새로운 데이터들이 추가되었을 때, 결과가 어떻게 변하는지를 미리 예측해볼 수 있게 한다.

<표 4-9> 단어 당 영상 크기에 따른 KU 데이터 세트 구성

데이터 세트	전체 영상	단어 당 영상	데이터 증강
KU 데이터 세트 10	410	10	24,600
KU 데이터 세트 15	615	15	36,900
KU 데이터 세트 28	1,148	28	68,880

<표 4-10> 다양한 크기의 KU 데이터 세트에 대한 실험 결과

데이터 세트	정확도
KU 데이터 세트 10	0.63
KU 데이터 세트 15	0.76
KU 데이터 세트 28	**0.87**
KU 데이터 세트 28 (2D CONV 제거)	0.85

<표 4-10>은 본 실험의 결과를 보여준다. 결과에서 볼 수 있듯이 본 논문에서의 실험은 매우 높은 단어 정확도를 보여준다. 또한 기존 3D 컨볼루션으로 구성되어 있던 네트워크에 2D 컨볼루션을 추가하여, 약 2%의 정확도 상승을 이뤄냈다. 이로 인해 본 논문에서 제안한 KU 한국 수어 데이터 세트와 단어 인식 네트워크는 87%의 높은 단어 인식률을 보여준다.

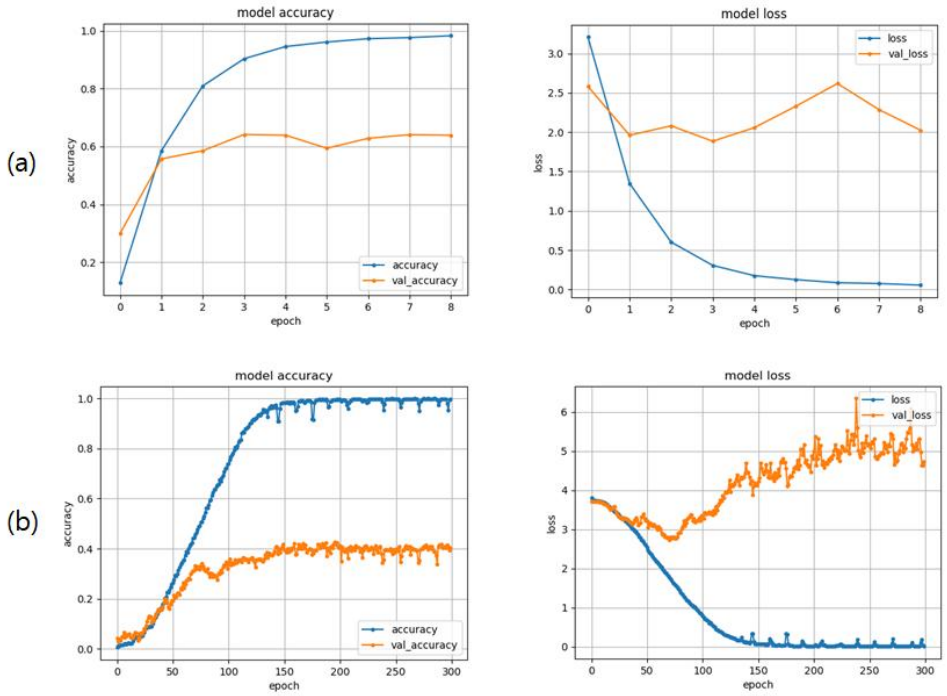
4. KU 수어 단어 인식 네트워크에서의 KETI 데이터 세트 실험 및 결과

본 논문에서 제안하는 KU 한국 수어 단어 인식 네트워크의 성능을 비교하기 위해 전자부품연구원에서 공개한 KETI 데이터 세트를 사용한다. 또한, 본 실험은 KETI 데이터 검증에 사용한 Seq2Seq 네트워크[14]를 적용하여, 본 논문에서 제안하는 네트워크와 비교한다. 본 실험에서는 KETI 데이터 세트와 KU 데이터 세트에서의 성능 비교를 위해 KETI 데이터 세트의 규모를 총 415단어에서 41단어를 갖는 작은 데이터 세트로 변경했다. 제안한 네트워크의 성능 평가를 위해 다음과 같은 4가지 방법의 실험을 진행한다.

- KU 수어 단어 인식 네트워크에서 KU 데이터 세트 사용
- KU 수어 단어 인식 네트워크에서 KETI 데이터 세트 사용
- KETI 네트워크에서 KU 데이터 세트 사용
- KETI 네트워크에서 KU 데이터 세트 사용

KU 수어 단어 인식 네트워크에서 사용된 입력 시퀀스의 구성은 100 x 100의 입력 크기와 30 프레임의 길이를 가진다. KETI 데이터 세트와 KU

데이터 세트 모두 이미지 자르기, 노이즈 추가, 랜덤 시퀀스 추출 이미지 증강 작업을 완료 후 사용하였다. KETI 네트워크에서는 Openpose로부터 추출한 키포인트를 입력 데이터로 사용한다. 이때 모든 키포인트를 사용하는 것은 아니며, 얼굴을 구성하는 키포인트를 제외한 137개의 키 포인트를 사용한다. 이미지 증강 작업은 프레임 영역을 분할 후 무작위로 프레임을 추출하는 랜덤 시퀀스 작업을 제외하면 별도의 이미지 증강은 진행하지 않는다. 두 데이터 세트 모두 훈련, 검증, 시험의 비율을 8 : 1 : 1로 분할하여 사용했다.



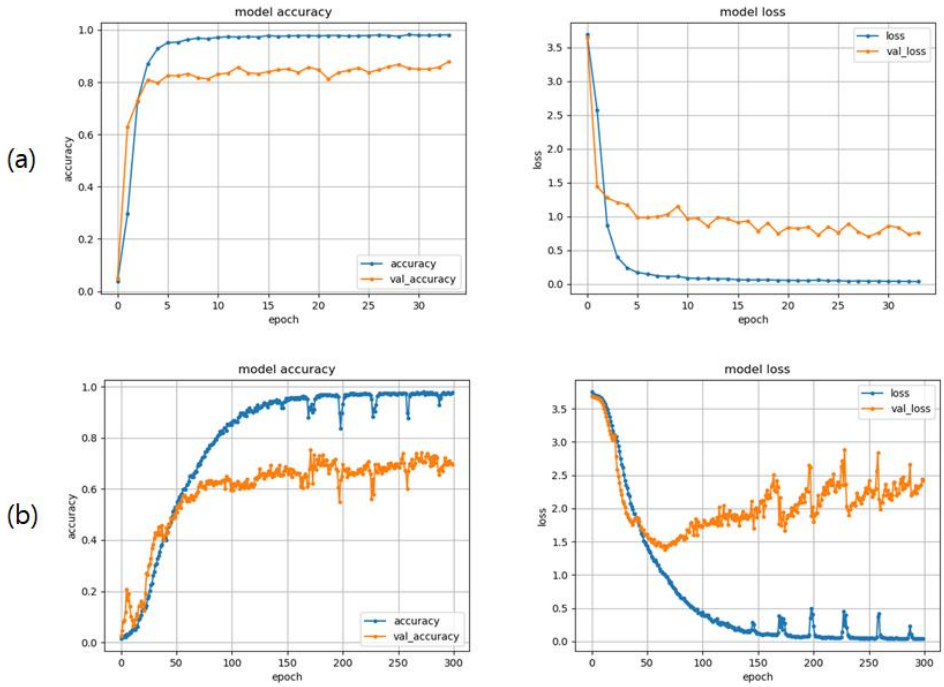
<그림 4-17> KU 데이터 세트를 사용한 네트워크 비교

(a) KU 수어 단어 인식 네트워크의 결과, (b) KETI 네트워크의 결과

<그림 4-17>은 KU 데이터 세트를 사용한 KU 수어 단어 인식 네트워

크와 KETI 네트워크를 비교한 실험 결과다. KU 수어 단어 인식 네트워크는 정확도와 손실값 그래프에서 KETI 네트워크보다 더 좋은 결과를 보여준다.

<그림 4-18>는 KETI 데이터 세트를 사용한 KU 수어 단어 인식 네트워크와 KETI 네트워크를 비교한 실험 결과다. KU 데이터 세트와는 다르게 두개의 네트워크 모두 준수한 결과를 보여준다.



<그림 4-18> KETI 데이터 세트를 사용한 네트워크 비교

(a) KU 수어 단어 인식 네트워크의 결과, (b) KETI 네트워크의 결과

결과적으로 KETI 데이터 세트와 KU 데이터 세트 모두 KU 수어 단어 인식 네트워크에서 더 좋은 결과를 보여줬다. 이 결과는 <표 4-11>에서도 볼 수 있다. <표 4-11>은 시험 데이터를 실험한 결과다. 실제 시험 데이터

를 사용한 결과에서도 KETI 네트워크와 비교할 때, KU 데이터 세트의 경우 33%p의 정확도 향상을, KETI 데이터 세트의 경우 22%p의 정확도가 향상된 것을 볼 수 있다. 즉, 두 데이터 세트 모두 KETI 네트워크보다 KU 수어 단어 인식 네트워크가 더 좋은 결과를 보여준 것을 볼 수 있다.

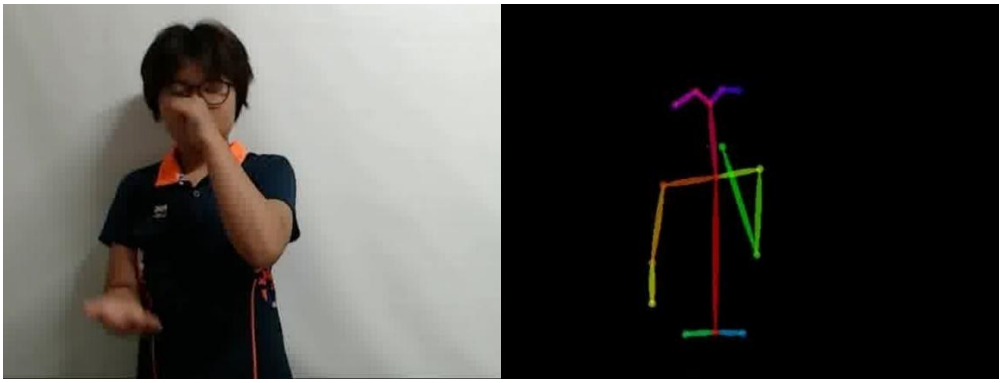
<표 4-11> 시험 데이터를 사용한 데이터 세트 비교 실험 결과

	KU 수어 단어 인식 네트워크		KETI 네트워크	
	정확도	**손실값**	**정확도**	**손실값**
KU 데이터 세트	0.76	1.39	0.43	4.40
KETI 데이터 세트	0.91	0.41	0.69	1.82

특기할 점은 단어당 영상의 수가 더 적은 KETI 데이터 세트가 더 높은 정확도를 보여준다는 것이다. KETI 데이터 세트가 더 높은 정확도를 보여주는 이유는 수어 영상 촬영 환경 때문이다. KETI 데이터 세트의 촬영은 전문 스튜디오에서 촬영한 영상이다. 단색의 깨끗한 배경, 밝은 조명 및 모든 촬영자들이 동일한 의상을 착용하고 전문 장비를 이용해 촬영이 진행됐다. 그에 비해 KU 데이터 세트는 복잡한 배경, 살짝 어두운 조명, 다양한 의상 등 일상생활 속 환경 내에서 일반적인 웹캠을 사용해 촬영이 진행됐다. 이러한 차이는 스켈레톤 영상을 만들 때 큰 차이를 보이게 된다. 복잡한 배경과 의상으로 인해 <그림4-19> 같이 스켈레톤 영상에서 정상적으로 손 또는 팔이 검출이 안되거나, 배경을 사람으로 인식하는 경우도 발생한다. 또한 카메라의 차이로 인해 <그림 4-20> 같이 영상내 잔상이 발생하게 되며,

정상적인 스켈레톤 영상의 추출이 힘들게 된다. 이런 잘못된 스켈레톤 영상의 추출은 정확도의 손실로 이어지게 된다. 이러한 이유 때문에 KU 데이터 세트의 실험 결과는 KETI 데이터 세트보다 낮은 정확도를 보여주지만, 실제 생활 환경에서는 더 높은 결과를 보여줄 것으로 보인다.

<그림 4-19> 배경으로 인한 오검출 스켈레톤 영상



<그림 4-20> 잔상으로 인한 오검출 스켈레톤 영상

제5장 결론 및 향후 연구

본 논문에서는 한국 수어 단어 인식을 위한 새로운 데이터 세트와 단어 인식 모델 네트워크를 제안했다. KU 한국 수어 데이터 세트는 포함하고 있는 단어의 수는 적지만, 전 세계 다른 수어 데이터 세트와 비교해도 가장 많은 단어 당 영상 수를 보여준다. 본 논문에서는 부족한 데이터를 보강하기 위한 3가지 데이터 증강 방법을 사용하였으며, 이에 대한 효과는 매우 뛰어난 것을 확인할 수 있다. 입력 영상으로 RGB 영상이 아닌 스켈레톤 영상을 사용해 수어 인식의 정확도를 증가시켰다. 또한 KU 수어 데이터 세트에 적합한 모델을 연구하기 위해 LRCN, C3D등의 다양한 7개의 네트워크의 구조를 실험하여 KU 수어 단어 인식 네트워크 모델을 제안했다.

KU 수어 단어 인식 네트워크 모델은 입력 영상을 제한하는 대신 정확도를 상승시켰으며, 네트워크에 2D 컨볼루션을 추가하여 정확도 상승을 이뤄냈다. 이로 인해 본 논문에서 제안한 KU 한국 수어 데이터 세트와 단어 인식 네트워크는 87%의 높은 단어 인식률을 보여준다. 이는 실생활에 적용하기에는 여전히 낮은 수치지만, 기존의 수어 인식 연구들의 인식률보다는 매우 높은 인식률을 보인다. KU 수어 단어 인식 네트워크모델의 성능을 평가하기 위해 KETI 수어 데이터 세트로 훈련 후 실험 한 결과 91%의 높은 정확도를 확인할 수 있었다.

본 연구는 매우 한정된 하드웨어 자원내에서 연구되었다. 그럼에도 불구하고 현재 발표된 한국 수어 연구에서 가장 높은 정확률을 보여준다. 이는 중간 실험 결과들이 보여주는 바와 같이 보다 많은 하드웨어 자원이 주어진다면 더 많은 프레임을 추출하여 적용할 수 있을 것이며, 이 경우 더 좋은

성능을 낼 수 있는 가능성을 나타낸다.

앞으로는 다양한 배경과 수어 사용자들의 더 많은 단어 수어 영상을 수집해야 할 것이며, 이를 통해 더 규모가 큰 KU 한국 수어 데이터 세트로 확장해야 할 것이다. 이는 장기적으로 수어 인식 연구를 진행할 많은 사람들에게 큰 도움이 될 것이다. KU 한국 수어 데이터 세트가 공개되면 좀더 많은 수어 연구가 진행되고 더 좋은 결과를 보여줄 것으로 기대된다.

또한 보다 높은 정확도를 위한 네트워크의 연구도 필요하다. 현재 수어 연구는 한정된 단어, 한정된 상황 등 제한된 범위 내에서만 연구가 진행되고 있다. 때문에 수어 인식 네트워크 정해진 범위 내에서만 연구가 진행되고 있다. 이를 위해 좀더 다양한 상황, 더 많은 단어들을 위한 네트워크의 연구가 진행되어야 한다. 또한 수어 인식을 위한 네트워크는 매우 많은 하드웨어 자원을 사용한다. 향후 연구를 통해 하드웨어 자원은 덜 소모하면서, 더 높은 성능을 낼 수 있는 수어 인식 네트워크의 연구가 필요하다.

참고문헌

[1] 이대섭, 「2017 한국수어사용실태조사연구 최종보고서」, 국립국어원, 2017.

[2]보건복지부,「장애인현황」, 2020년 04월, http://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_11761_N005

[3] 양승한, 정승준, 강희광, 김창익,「한국 수화 인식을 위한 데이터셋」, 『대한전자공학회 학술대회』, 480-488, 2019.6.

[4] Erdem Yörük, Ender Konukoglu, Bülent Sankur, Jérôme Darbon, 「Shape-based hand recognition」, IEEE transactions on image processing 』, VOL.15, NO.7, pp. 1803-1815, 2006.

[5] Maria Eugenia Cabrera, Juan Manuel Bogado, Leonardo Fermin, Raul Acuna, Dimitar Ralev,「Glove-Based Gesture Recognition System」,『Adaptive Mobile Robotics』, pp. 747-753, 2012.

[6] Martinez, Aleix M., et al.,「Purdue RVL-SLL ASL database for automatic recognition of American Sign Language」,『Proceedings of IEEE International Conference on Multimodal Interfaces』, pp. 167-172, 2002.

[7] Carol Neidle, Ashwin Thangali, and Stan Sclaroff,「Challenges in development of the american sign language lexicon video dataset(asllvd) corpus 」,『5th Workshop on the Representaion and Processing of Sign Languages: Interactions between Corpus and Lexicon, Language Resources and Evaluation onference』, 2012.

[8] Lu, Pengfei, Matt Huenerfauth,「Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation」,『Computer Speech & Language』, pp. 812-831, 2014.

[9] Forster, Jens, et al,「RWTH-PHOENIX-Weather: A Large Vocabulary Sign language Recognition and Translation Corpus」, 『Proceedings of the Eighth International Conference on Language

Resources and Evaluation』, pp. 3785-3789, 2012.
[10] Von Agris, ULrich, et al, 「Recent developments in visual sign language recognition」, 『Universal Access in the Information Society』, pp. 323-362, 2008.
[11] Chai, X., et al, 「DEVISIGN: Dataset and Evaluation for 3D Sign Language Recognition」, Beijing Tech. Rep, 2015.
[12] Oszust, Mariusz, and Marian Wysocki, 「Polish sign language words recognition with kinect」, 『2013 6th InternationalConference on Human System Interactions 』, pp. 219-226, 2013.
[13] Kapuscinski, Tomasz, et al. 「Recognition of hand gestures observed by depth cameras」, 『International Journal of Advanced Robotic System』, pp. 4-26, 2015.
[14] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung. 「Neural Sign Language Translation based on Human Keypoint Estimation」, 『Applied Sciences』,9(13), 2019
[15] T. Starner, A. Pentland, 「Real-time American sign language recognition from video using hidden markov models」, 『Proceedings of International Symposium on Computer Vision - ISCV』, pp. 265-270, 1995.
[16] C. Dong, M. C. Leu, and Z. Yin., 「American sign language alphabet recognition using Microsoft Kinect」, 『Proceedings of IEEE Conference on Computer Vision and Pattern Recognition』, pages 44-52, 2015.
[17] S. Gattupalli, A. Ghaderi, and V. Athitsos., 「Evaluation of deep learning based pose estimation for sign language recognition」, 『Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments』, pp. 1-7, 2016.
[18] P. V. V. Kishore, A. S. C. S. Sastry, and A. Kartheek., 「Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds」, 『2014 First International Conference

on Networks Soft Computing』, pp. 135-140, 2014.
[19] O. Koller, J. Forster, and H. Ney.,「Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers」,『Computer Vision and Image Understanding』, 141:108-125, 2015.
[20] O. Koller, S. Zargaran, and H. Ney.,「Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms」,『Proceedings of IEEE Conference on Computer Vision and Pattern Recognition』, pp. 3416-3424, 2017.
[21] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, Richard Bowden,「Neural Sign Language Translation」,『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』, pp. 7784-7793, 2018.
[22] Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss,「The Significance of Facial Features for Automatic Sign Language Recognition」,『2008 8th IEEE International Conference on Automatic Face & Gesture Recognition』, pp. 1-6, 2008.
[23] Oscar Koller, Necati Cihan Camgoz, Hermann Ney, Richard Bowden,「Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos」,『IEEE transactions on pattern analysis and machine intelligence』, 2019.
[24] Koustav Mullick, Anoop M. Namboodiri,「 Learning deep and compact models for gesture recognition 」,『IEEE International Conference on Image Processing』, pp. 3998-4002, 2017.
[25] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, Benjamin Schrauwen,「Sign Language Recognition Using Convolutional Neural Networks,『European Conference on Computer Vision, pp. 572-578, 2015.
[26] K. Simonyan and A. Zisserman,「Two-stream convolutional

networks for action recognition in videos」, 『Advances in neural information processing systems』, pp. 568−576, 2014.
[27] H. Liu, J. Tu, and M. Liu, 「Two−stream 3d convolutional neural network for skeleton−based action recognition」, abs/1705.08106, 2017.
[28] A. Krizhevsky, I. Sutskever, G. Hinton, 「ImageNet Classification with Deep Convolutional Neural Networks」, 『Advances in neural information processing systems』, pp. 1097−1105, 2012.
[29] K. Simonyan, A. Zisserman, 「Very deep convolutional networks for large−scale image recognition」, arXiv:1409.1556, 2014.
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, A. Rabinovich, 「Going deeper with convolutions」, 『Proceedings of the IEEE conference on computer vision and pattern recognition』, pp. 1−9, 2015.
[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, 「Deep Residual Learning for Image Recognition」, 『 Proceedings of the IEEE conference on computer vision and pattern recognition』, pp. 770−778, 2016.
[32] S. Hochreiter, J. Schmidhuber, 「Long short−term memory」, 『Neural Computation』, 9(8), pp. 1735−1780, 1997.
[33] Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu, 「3D Convolutional Neural Networks for Human Action Recognition」, 『IEEE transactions on pattern analysis and machine intelligence』, 35(1), pp. 221−231, 2013.
[34] Ken−ichi Funahashi, Yuichi Nakamura, 「Approximation of dynamical systems by continuous time recurrent neural networks」, 『Neural Networks』, 6(6), pp. 801−806, 1993.
[35] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. Barbano, 「Toward Automatic Phenotyping of Developing Embryos

from Videos」，『IEEE Transactions on Image Processing』，14(9)，pp. 1360−1371, 2005.
[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri,「Learning spatiotemporal features with 3D convolutional networks」，『Proceedings of the IEEE international conference on computer vision』，pp. 4489−4497, 2015.
[37] Dongxu Li , Cristian Rodriguez Opazo, Xin Yu, Hongdong Li,「Word−level Deep Sign Language Recognition from Video: A New Large−scale Dataset and Methods Comparison」，『The IEEE Winter Conference on Applications of Computer Vision』，pp. 1459−1469, 2020.
[38] Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman,「Convolutional Two−Stream Network Fusion for Video Action Recognition」，『Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition』，pp. 1933−1941, 2016.
[39] C. McCaskill, C. Lucas, R. Bayley, J. Hill,「The hidden treasure of Black ASL: Its history and structure」，『Gallaudet University Press Washington』，DC, 2011.
[40] 「The 20bn−jester dataset−v1」,2020−05−10, https://20bn.com/datasets/jester.
[41] Amir, Arnon, et al.,「A low power, fully event−based gesture recognition system」，『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』，pp. 7243−7252, 2017.
[42] K. Soomro, A. R. Zamir, M. Shah, 「Ucf101: A datasetof 101 human actions classes from videos in the wild」，arXivpreprint arXiv:1212.0402, 2012.
[43] J. Carreira and A. Zisserman. 「Quo vadis, action recognition?a new model and the kinetics dataset」，『Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition』，pp. 6299−6308, 2017.

[44] 「Asl university」, http://asluniversity.com/
[45] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, K. Emmorey, 「Asl-lex: A lexical database of american signlanguage」, 『Behavior research methods』, 49(2), pp. 784-801, 2017
[46] A. Karpathy, G. Toderici, S.Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, 「Large-scale video classification with convolutional neural networks」, 『Proceedings of the IEEE conference on Computer Vision and Pattern Recognition』, pp. 1725-1732, 2014.
[47] J. Donahue, et al., 「Long-term recurrent convolutional networks for visual recognition and description」, 『Proceedings of the IEEE conference on computer vision and pattern recognition』, pp. 2625-2634, 2015.
[48] X. Ouyang, S. Xu, C. Zhang, P. Zhou, Y. Yang, G. Liu, X. Li, 「A 3D CNN and LSTM Based Multi-Task Learning Architecture for Action Recognition」, 『IEEE Access』, 7, pp. 40757-40770, 2019
[49] B. Mahasseni, S.Todorovic, 「Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition」, 『Proceedings of the IEEE conference on computer vision and pattern recognition』, pp. 3054-3062, 2016.
[50] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, 「Beyond short snippets: Deep networks for video classification」, 『Proceedings of the IEEE conference on computer vision and pattern recognition』, pp. 4694-4702, 2015.
[51] S. Escalera, X. Bar´o, J. Gonz alez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-L´opez, H. J. Escalante, J. Shotton, I. Guyon, 「ChaLearn Looking at People Challenge 2014: Dataset and Results」, 『European Conference on Computer Vision』, pp. 459-473, 2014.
[52] Z. Cao, T. Simon, S. Wei, Y. Sheikh, 「Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields」, 『Proceedings of the

2017 IEEE Conference on Computer Vision and Pattern Recognition』, pp. 1302-1310, 2017.
[53] T. Simon, H. Joo, I.A. Matthews, Y. Sheikh, 「Hand Keypoint Detection in Single Images Using Multiview Bootstrapping」, 『In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition』, pp. 4645-4653, 2017.
[54] S. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, 「Convolutional Pose Machines」, 『Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition』, pp. 4724-4732, 2016.
[55] M. Zahedi, D. Keysers, T. Deselaers, H. Ney, 「Combination of tangent distance and an image distortion model for appearance-based sign language recognition」, 『Joint Pattern Recognition Symposium』, pp. 401-408, 2005.
[56] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, 「Temporal segment networks: Towards good practices for deep action recognition」, 『European conference on computer vision』, pp. 20-36, 2016.
[57] N. Pinto, D. D. Cox and J. J. DiCarlo, 「Why is real-world visual object recognition hard? 」, 『PLoS Computational Biology 2008.
[58] Nair, V., Hinton, G.E.: 「Rectified linear units improve restrcted Boltzmann machines」, 『Proceedings of the 27th International Conference on Machine Learning』, pp. 807-814, 2010.
[59] R. Poppe, 「A survey on vision-based human action recognition. Image and Vision」, 『Computing 28』, pp. 976-990, 2010.
[60] K. Cho, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation", arXiv:1406.1078, 2014.
[61] Y. Yang and D. Ramanan. 「Articulated pose estimation with flexible mixtures-of-parts」, 『CVPR 2011』, pp. 1385-1392, 2011.

[62] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. 「Poselet conditioned pictorial structures」, 『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』,pp. 588-595, 2013.
[63] A. Toshev, C. Szegedy, and G. 「DeepPose: Human pose estimation via deep neural networks」, 『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』, pp. 24-27, 2014.
[64] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. 「Human pose estimation with iterative error feedback」, 『Proceedings of the IEEE conference on computer vision and pattern recognition』,pp. 4733-4742, 2016.
[65] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. 「Multi-context attention for human pose estimation」, 『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』, pp. 1831-1840, 2017.
[66] X. Chu, W. Ouyang, H. Li, and X. Wang. 「Structured feature learning for pose estimation」, 『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』, pp. 4715-4723, 2016.
[67] W. Yang, W. Ouyang, H. Li, and X. Wang. 「End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation」, 『Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition』, pp. 3073-3082, 2016.
[68] Chiu, H. K., Adeli, E., Wang, B., Huang, D. A., & Niebles, J. C. 「Action-agnostic human pose forecasting」, 『2019 IEEE Winter Conference on Applications of Computer Vision』, pp. 1423-1432, 2019.

부 록

KU 한국 수어 데이터 세트

①	②	③	④
⑤	⑥	⑦	⑧

<KU 한국 수어 데이터 세트 이미지 순서>

상기 표는 부록에 수록된 KU 데이터 세트의 이미지를 보는 순서를 보여준다. 표에 표시된 번호 순대로 왼쪽 상단부터 우측으로 진행되며, 그후 왼쪽 하단부터 우측으로 진행된다.

<사죄>

<빌다>

<조심>

<연락>

<멋지다>

<귀엽다>

<싫어하다>

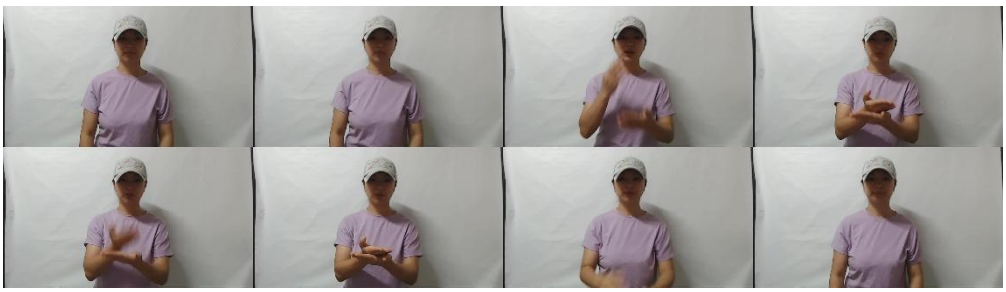
<수고>

<부끄럽다>

<매우>

<앞으로>

<즐겁다>



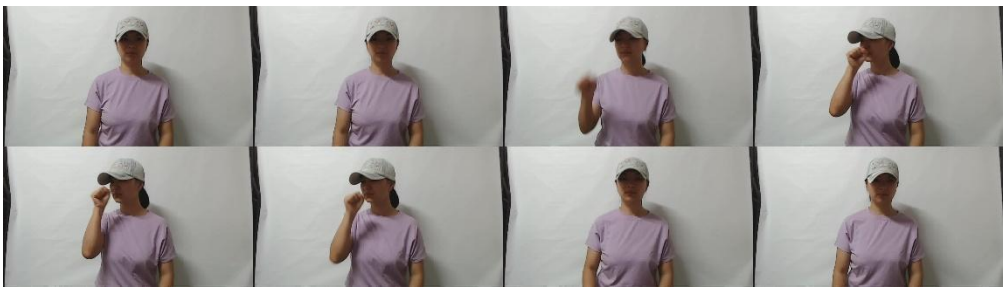
<주십시오>

<좋다>

<안녕하세요>

<친절>

<늦다>



<좋아하다>

<만나다>

<실수>

<절대로>

<반갑다>

<안되다>

<의견>

<예쁘다>

<질문>

<정말>

<화해>

<안타깝다>

<부탁>

<맞다>

<똑같다>

<미안>

<죄송하다>

<덕분>

<고맙다>

<못생기다>

<용기>

<기다리다>

<천만에>

<말씀>

국문초록

심층 신경망을 이용한 한국 수어 단어 인식

배효철

컴퓨터 · 정보통신공학과

건국대학교 대학원

수어는 청각 장애인들이 일상어로 사용하는 하나의 언어이다. 하지만 청각 장애인들이 수어를 이용해 일반인들과 의사소통을 하기에는 매우 어렵다. 이러한 문제들을 해결하기위해 세계의 여러 나라에서는 이미 수어를 위한 다양한 연구가 진행되어 왔다. 하지만 한국은 수어 연구에 많이 뒤쳐져 있다. 그로 인해 한국은 수어 연구에 필요한 데이터 세트가 전무한 실정이다. 수어 연구에 있어 수어 데이터 세트는 매우 중요하다.

본 논문에서는 한국 수어의 인식을 위한 새로운 데이터 세트와 딥러닝 네트워크 모델을 제안하여 이러한 문제들을 해결하고자 한다. 본 논문에서 제안하는 KU 한국 수어 데이터 세트는 일상 생활에서 많이 사용되는 인사 및 대화를 주제로 삼는 41개의 단어를 총 10명의 일반인이 녹화한 총 1,151개의 영상으로 구성된다. KU 한국 수어 데이터 세트는 포함하고 있는 단어의 수는 적지만, 전 세계 다른 수어 데이터 세트와 비교해도 가장 많은 수를 보여준다. 또한 부족한 데이터를 보강하기 위한 3가지 데이터 증강 방법을 사용하였으며, 이에 대한 효과는 매우 뛰어난 것을 보여준다.

추가로 본 논문에서는 한국 수어 단어 인식을 위한 단어 인식 모델 네트워크를 제안한다. 스켈레톤 영상을 사용해 수어 인식의 정확도를 증가시켰으며, 단어 인식 네트워크에 2D 컨볼루션을 추가하여, 약 2%의 정확도 상승을 이뤄냈다. 이로 인해 본 논문에서 제안한 KU 한국 수어 데이터 세트와 단어 인식 네트워크는 87%의 높은 단어 인식률을 보여준다.

본 논문의 구성은 다음과 같다. 2장에서는 수어 연구를 위한 배경지식과 관련된 연구인 여러 국가들의 다양한 데이터 세트 및 기존의 동작인식 및

수어 연구에 사용된 기술들에 대해 살펴본다. 3장은 본 논문에서 제안하는 한국 수어를 위한 데이터 세트에 대한 설명과 새로운 한국 수어 인식을 위한 딥러닝 네트워크를 제안한다. 4장에서는 입력 영상의 크기, 입력 영상의 길이, 데이터 증강 및 네트워크 구성에 대한 다양한 방식을 실험해보고 그에 대한 측정 결과에 대해 기술한다. 마지막으로 5장에서는 앞으로 진행되어야 할 한국 수어 인식 연구에 대해 제안한다.

주제어(9P) : 3D CNN, 수어, 수화, 데이터 증강, 딥러닝