https://doi.org/10.7840/kics.2023.48.6.669

관절 좌표 이미지 패치를 이용한 다중 스트림 신경망을 통한 수화 인식

강 현 석◆, 박 광 현°

Sign Language Recognition with Multi-Stream Neural Network Using Joint Point Image Patches

Hyeon Seok Kang◆, Kwang-Hyun Park°

요 약

최근 들어 딥러닝(Deep Learning)과 기계학습(Machine Learning) 알고리즘을 이용한 수화 인식에 관한 연구가 많이 진행되고 있다. 다양한 종류의 정보를 요구하는 수화의 특성 때문에 단어 단위 수화 인식은 다양한 방법론들이 제시되었다. 수화 인식에 대한 기법은 이미지 기반의 방법과 자세 기반의 방법으로 나뉜다. 초반에는 이미지 기반의 합성곱 신경망을 이용하는 방법들이 나와서 준수한 성능을 보였다. 이미지 기반의 방법들은 주로 수화의 전체적인 정보에 집중하였다. 그 이후에는 행동 인식 분야에서 많이 사용하던 자세 기반의 관절 정보를 이용한 그래프 합성곱 신경망을 적용하였다. 이때, 얻은 정보들은 대부분 수화에서 관계성에만 집중하였다, 하지만, 단일 스트림 신경망으로 얻을 수 있는 정보로는 다양한 종류의 정보가 필요한 수화 특징을 잡아내는데 부족했고, 많은 논문이 다중 신경망을 통해서 수화 인식에 필요한 다양한 정보를 얻었다. 본 논문에서는 수화 인식에서 중요한 얼굴과 손의 지역적인 정보를 얻기 위한 데이터 형태인 관절 좌표 이미지 패치와 트랜스포머 네트워크를 활용한 다중 스트림 네트워크를 제안한다.

키워드 : 수화 인식, 다중 스트림 신경망, 관절 좌표 이미지 패치
Key Words : Sign Language Recognition, Multi-stream Neural Network, Joint Point Image Patch

ABSTRACT

Recently, many studies on sign language recognition using deep learning and machine learning algorithms have been conducted. Due to the nature of sign language requiring various types of information, various methods have been proposed for word-level sign language recognition. Sign language recognition techniques are divided into image-based methods and pose-based methods. In the beginning, methods using image-based convolutional neural networks came out and showed satisfactory performance. Image-based methods mainly focused on the overall information of sign language. After that, a graph convolutional neural network using pose-based joint information, which was widely used in the field of action recognition, was applied. At this time, most of the information obtained was focused only on relationships in sign language. However, the information obtained with a single-stream neural network was insufficient to capture sign language features that

※ 이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0017124, 2023년 산업혁신인재성장지원사업)
◆ First Author : KwangWoon University, School of Robotics, deimon213@naver.com, 학생회원
° Corresponding Author : KwangWoon University, School of Robotics, akaii@kw.ac.kr, 정회원
논문번호 : KICS 202301-005-B-RN, Received January 10, 2023; Revised March 6, 2023; Accepted March 15, 2023

required various types of information, and many papers have been published on sign language recognition through multi-stream neural networks. In this paper, we propose a multi-stream network using joint point image patches, which are data types to obtain face and hand regional information, and a transformer network.

Ⅰ. 서 론

수화는 농인들을 위해 만들어진 언어이다. 하나의 완전한 언어로서 문법과 단어들이 일반적인 언어와는 다르기 때문에 배우고 사용하는 데 큰 어려움이 있다. 또한, 수화는 표정, 손의 이동 정도, 손의 모양에 따라 단어의 뜻이 달라진다. 일반적인 언어에서 사람마다 억양 차이가 있듯이 수화도 사람마다 정도의 차이가 있어서 같은 단어도 다르게 표현되는 경우가 있다. 이러한 다양성 때문에 수화 인식에는 여러 종류의 정보가 필요하고, 컴퓨터 비전 분야에서 큰 관심을 받았다. 수화 인식은 크게 문장 단위와 단어 단위의 인식으로 나뉘는데, 전자는 문장 전체를 한 번에 다루면서 문장 단위의 해석 및 이해에 집중하고, 후자는 단어 자체의 구분과 인식에 집중한다. 본 논문은 단어 단위의 수화 인식을 다룬다.

최근 딥러닝과 트랜스포머 모델의 등장으로 다양한 방법들이 제안되었다. 수화 인식 분야 또한 여러 방법론들이 제안되었다, 그 중에는 다른 분야에서 최고의 성능을 보인 네트워크도 있는데, 수화 인식에서는 그만큼의 좋은 성능을 보이지 못했다. 그 이유로는 첫째, 앞서 언급했듯이 수화는 사람의 표정, 손의 모양, 손의 움직임 등 다양한 정보를 사용하여 구분된다. 따라서, 수화를 완벽하게 구분하기 위해서는 손에 관한 다양한 정보가 필요하다. 두 번째로 영상에서 얻어내는 프레임 자체의 문제들이 있다. 일부 수화자들의 수화 영상을 보면 손이 갑자기 빠르게 움직이거나 다른 손에 의해 가려지는 등의 문제로 손 자체를 인식하는 데 문제가 발생한다. 이러한 경우에는 문제가 생긴 프레임을 제외하고 학습을 진행하거나, 이전 혹은 이후의 프레임을 통해 정보를 채우는 방식으로 손실된 정보를 보충할 수 있다. 마지막으로 일반적인 언어에 동음이의어 혹은 이음동의어가 존재하고 사람마다 말하는 억양이나 습관이 다르듯이 수화 역시 그러한 단어들이 많이 존재한다. 그림 1은 수화가 유사하지만 의미가 다른 예시들을 보여준다. 수화의 경우에는 단어들 간의 차이점이 수화자의 입 모양, 손의 모양, 움직임 정도가 있는데, 수화 영상에서는 크게 차이가 나지 않는다. 따라서 단순하게 전체적인 정보만 학습하는 게 아니라 지역적인 정보를 학습하는 것도 필요하다. 또

그림 1. 유사하지만 다른 수화 예시 (1: 소망, 2: 배고픔, 3: 밥, 4: 스프) [1]
Fig. 1. Examples of similar but different sign language (1: Wish, 2: Hungry, 3: Rice, 4: Soup)

한, 공간적인 정보 뿐만 아니라 시간적인 정보도 중요하게 다루어야 한다.

얼굴 인식 분야에서는 얼굴을 구분하기 위해 얼굴 전체보다는 특징점인 눈, 입, 코 같은 지역적인 부분에 집중하는 방법론이 제시되었다[2]. 수화 인식에서도 지역적인 정보와 전체적인 정보를 결합해서 이용하는 다중 네트워크 방법이 등장하고 있다[3]. 지역적인 정보에 집중하기는 했지만, 지역적인 정보만 다루지 않고 손과 얼굴의 전체적인 외형에 대한 정보도 다루었다. 하지만 위치가 잘 변하지 않는 얼굴과는 다르게 손은 계속 위치가 바뀌는데 단순하게 손 주변의 이미지를 사용하면 현재 그 손이 어디에 있는지 알 수 없다. 본 논문은 이러한 기법들[2,3]에서 영감을 받아 수화자의 얼굴 및 손의 지역적인 정보와 위치를 고려한 관절 좌표 이미지 패치, 정보들의 관계성을 학습하기에 유리한 트랜스포머 네트워크를 기반으로 하는 다중 스트림 네트워크를 제안한다. 제안하는 네트워크의 전체적인 구조는 그림 2와 같다. 각 스트림에 대한 자세한 내용은 Ⅱ절에서 자세하게 다룬다.

앞서 언급했듯이 수화는 크게 두 가지로 나뉘는데, 연속적인 수화 인식과 분리된 수화 인식이다. 전자는 문장 단위의 인식으로서 단어 하나의 이해보다는 문장 전체를 이해하는 데 집중한다. 즉, 단어를 따로 분

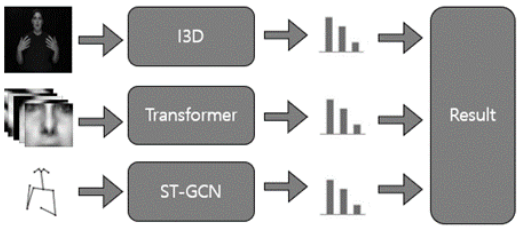


그림 2. 전체적인 네트워크 구조
Fig. 2. Structure of whole network

리하는 것이 아니라 문장 전체를 하나로 보고 구분한다. 반면에, 후자는 단어 단위의 수화 인식으로서 단어 단위로 쪼개진 수화를 인식하는 데 집중한다. 일반적으로 연속적인 수화 인식이 문장을 한 번에 분석하고 해석하기 때문에 더 어려운 문제이다.

전통적인 수화 인식은 손짓 인식과 유사하게 손 인식용 장갑 형태의 장비를 착용하고 행동 인식 센서를 통해 손 모양을 인식하는 방법을 사용했다. 또한 행동 인식 분야처럼 수화를 하나의 패턴으로 보고 확률론적인 관점에서 수화를 인식하고자 하였다. 최근에는 3차원 인공신경망과 트랜스포머의 도입으로 시간 영역의 정보를 얻을 수 있게 되면서 수화 인식 분야에서도 성능이 많이 개선되었다. 수화 인식의 방법론은 이미지 기반[1,4,5]와 자세 기반[1,6,7]의 방법론으로 나뉜다.

이미지 기반의 방법론들은 대부분 공간적인 정보에 집중하고, 수화자의 모습 같은 시각적인 정보를 위주로 다루었다. 이미지 기반의 가장 대표적인 방법론은 Inflated Inception 3D (I3D)[4]가 있다. 네트워크 구조는 그림 4에서 볼 수 있다. 행동 인식 분야에서 좋은 성능을 보인 네트워크가 수화 인식에서도 좋은 성능을 보였다. 하지만, 수화자의 상체와 손 정보, 그 외 다른 인식에 불필요한 정보를 모두 일괄적으로 학습하는 특징 때문에 한계점이 명확했다. 이러한 문제를 해결하기 위해 트랜스포머 중에서 계층 구조를 활용하여 객체 탐지에 좋은 성능을 보인 Swin Transformer[5]를 수화 인식에 적용하여 필요 없는 주변 정보의 학습량을 낮추고 중요한 정보가 있는 사람의 상체부분에 집중함으로써 성능을 개선하였다.

자세 기반의 방법론은 행동 인식에서 많이 사용되었던 스켈레톤을 이용해서 관절 자체의 정보를 그래프로 만들어 학습하는 방법이다. 그림 3은 자세 기반의 방법론에 대한 예시를 보여준다. 기본적인 방법론으로는 그래프 합성곱 신경망(GCN)이 있으며, 이미지 기반과는 다르게 시계열 정보에 집중한다. 한 프레임의 정보가 아니라 모든 프레임에 걸쳐 관절 좌표 간

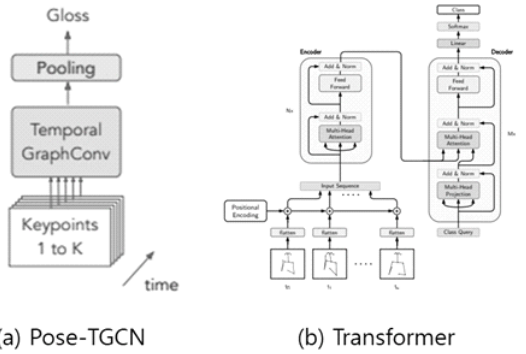


그림 3. 자세 기반 방법의 예시 [1,7]
Fig. 3. Examples of pose-based methods

의 관계성에 집중하여 학습한다. 자세 기반의 방법론에서 가장 일반적인 예시는 3차원 합성곱 신경망을 사용한 Pose-TGCN[1]이 있다. 단순하게 수화 영상에서 프레임마다 주요 포인트인 관절 좌표를 그래프로 만들어서 학습하는 방식이다. 초기에는 I3D처럼 괜찮은 성능을 보였으나 수화의 복잡한 특성 때문에 금방 한계를 보였다. 이미지 기반의 방법론에서 사용한 트랜스포머와는 다르게 자세 기반의 방법론은 자연어 처리에서 사용하는 BERT[6]와 인코더-디코더 형태의 트랜스포머[7]를 사용했다. 관절 좌표를 이용해서 그래프 혹은 스켈레톤 형태를 만들어서 입력으로 활용하는데 그림 3(b)에서 확인할 수 있다. 이러한 방식 또한 단순하게 시간적인 측면에서 전체적인 정보만을 학습하기 때문에 좋은 성능을 보이지는 못했다.

Ⅱ. 제안하는 네트워크

2.1 다중 스트림 네트워크

수화 인식 분야는 어느 특정한 성질의 정보만으로는 구분하는 데 한계가 명확하다. 얼굴 인식 연구[2]와 수화 인식 연구[3]에서 영감을 받아 수화 이미지의 전체적인 정보와 지역적인 세부 정보를 결합하기 위해 다중 스트림 네트워크를 이용한다. 각 입력 데이터는 각자의 네트워크를 통과하고 정확도 점수를 앙상블 기법으로 합쳐서 최종적인 정확도를 산출한다. 다중 스트림 네트워크는 총 3가지의 네트워크로 구성되어 있는데 전체적인 네트워크 구조는 그림 2에서 확인할 수 있다. 첫 번째 스트림은 기본 스트림으로서, 수화 영상에서 수화자의 외형 및 전체적인 정보를 학습하기 위해 사용한다. 두 번째 스트림은 관절 좌표 이미지 패치 스트림으로서, 관절 좌표 주변으로 이미지 패치를 만들어 수화 영상에서 지역적인 정보를 얻기 위

해 사용한다. 마지막 스트림은 스켈레톤 스트림으로서 자세 기반의 방법론으로 손, 얼굴과 상체의 위치에 대한 관계성을 학습하기 위해 사용한다.

2.2 기본 스트림

수화자의 외형과 전체적인 정보를 학습하기 위한 스트림으로서 단순하게 이미지만 이용해서 학습한다. 학습 모델로는 I3D를 사용하는데 수화 인식에서 좋은 성능을 보였기 때문이다. I3D는 그림 4에서 전체적인 흐름을 확인할 수 있다. 이 네트워크는 ImageNet[8]으로 학습하고 Kinetics 데이터 세트[4]로 미세 조정을 한 후에 사용하는데, 영상의 공간적, 시간적인 정보를 더 잘 얻을 수 있게 된다. 네트워크의 전체적인 설정들은 [1]의 조건에 맞추어 설정하였다. 최종적으로 얻은 점수를 마지막 앙상블 층으로 보내어 인식 과정을 거치게 된다.

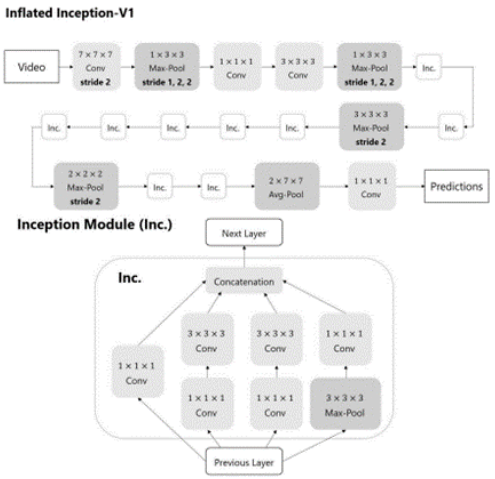


그림 4. I3D 네트워크 [1]
Fig. 4. I3D network

2.3 관절 좌표 이미지 패치 스트림

두 번째 스트림은 다른 스트림에서 얻지 못하는 수화의 지역적인 정보를 얻기 위한 관절 좌표 이미지 패치 스트림이다. 손의 동작은 같지만 손의 형태가 달라서 뜻이 달라지거나 다른 모든 것은 같지만 손의 위치에 따라서 뜻이 갈리는 수화가 존재하기 때문에 손의 모양과 위치는 수화 인식에서 중요한 요소 중 하나이다. 또한, 농인이 청인과 소통할 때 손만이 아니라 얼굴과 입 모양을 보고 소통하는 경우가 많아서 얼굴에서 입 주변의 정보도 중요한 부분이다. 이러한 상황에 맞추어 기존의 방법[3]은 손과 얼굴에 집중하기 위해 손과 얼굴 주변의 이미지를 따로 얻어서 각각 하나의 스트림으로 만들고, 총 3개의 스트림을 사용한다. 이러한 방식으로 어느 정도 지역적인 정보도 얻고 외형적인 정보도 학습에 활용할 수 있다. 하지만 일부 수화의 경우에는 좀 더 세부적인 정보를 요구한다. 예를 들어 손가락만 약간 움직여 표현하는 수화도 존재한다. 이러한 경우에는 단순하게 손 주변의 이미지만 얻게 되면 거의 다 같은 장면이 나와서 구분할 수 없다. 때문에 더 지역적인 정보와 손의 위치에 대한 정보를 학습할 필요가 있다. 이러한 정보를 학습에 활용하기 위해 49개의 관절 좌표를 기준으로 일정한 크기의 이미지 패치를 얻어서 학습에 활용한다.

우선, 스트림에서 사용하기 위한 손과 얼굴의 49개 좌표를 구하기 위해 구글의 MediaPipe를 사용하여 각각의 좌표 값들을 얻는다. 사람의 신체 전체에 대한 관절 좌표를 얻을 수 있는데, 사람의 상반신에 해당하는 15개의 좌표 중에서 일부인 코, 눈, 귀, 입에 해당되는 좌표 7개만 선택해서 추출한다. 또한, 각각의 손에서 21개의 좌표들을 선택해서 총 42개의 좌표를 얻고, 앞서 구한 7개의 좌표와 합쳐서 최종적으로 49개의 좌표를 얻는다. 그림 5는 49개 좌표의 예시를 보여준다. 이렇게 얻은 49개의 좌표는 일정한 순서를 가지고 식 (1)처럼 정리된다. f는 얼굴, h는 손을 의미하고, 아래 첨자는 좌표의 인덱스를 나타낸다. 얼굴에 7개, 오른손 21개, 왼손 21개로 구성된다. 이 때, 오른손과 왼손의 순서는 기본적으로 오른손 다음에 왼손 순서로 되어있지만 일부 좌표가 찍히지 않거나 한 손 수화인 경우에는 순서가 바뀌는 경우가 있다.

$$P_{skel} = \left(p_0^f, \ldots, p_6^f, p_0^{h1}, \ldots, p_{20}^{h2}, p_0^{h1}, \ldots, p_{20}^{h2}\right) \quad (1)$$

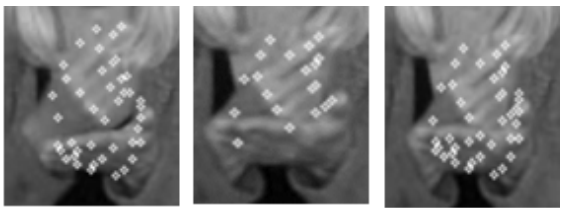
그림 5. 49개 좌표 이미지의 예시
Fig. 5. Example of 49 joint point image

이 좌표들을 기준으로 이미지 패치를 만든다. 하지만, 패치 추출 이전에 자세 기반의 수화 인식에서 생기는 고질적인 문제를 고려해야 한다. 손의 움직임이나 가려짐에 의해 손의 좌표가 정확히 찍히지 않거나 아예 찍히지 않는 문제이다. 이러한 문제를 해결하기 위해, 필요한 위치에 좌표가 찍히지 않은 경우에는 이전의 장면에 찍혀 있는 정보와 현재 장면에서의 좌표를 비교해서 이번 좌표를 추가로 채운다. 좌표를 채우기 위한 기준점은 얼굴은 코, 양손의 경우는 각각의 손목 좌표를 기준으로 한다. 이는 식 (2)에 간단하게 표현되어 있다. 여기서 P는 패치를 위한 좌표를 의미하고, T는 시간을 나타낸다. dis는 현재 프레임에 제대로 찍혀 있는 손목의 좌표와 이전 프레임에 찍혀 있는 다른 두 손목과의 거리를 의미한다. rh와 lh는 각각의 손목을 의미한다. 직접적으로 거리를 비교해서 거리가 더 멀리 있는 손목에 해당되는 값과 그 이후 20개의 좌표를 가지고 와서 비워진 자리를 채운다.

$$P_i^T, \ldots = P_{lh}^{T-1}, \ldots \quad if \;\; dis_{rh} < dis_{lh}, \\ else \;\; P_{rh}^{T-1}, \ldots if \;\; dis_{rh} > dis_{lh} \tag{2}$$

간단한 예시는 그림 6에서 확인할 수 있다. 이전 장면에서는 양손이 다 나오지만, 현재 장면에서는 한 손의 좌표만 나오게 되는데, 현재 장면에서 찍힌 한 손의 손목 좌표와 이전 장면에서 나온 양손의 손목 좌표 사이의 거리를 이용해서 새로운 좌표를 채운다.

각 좌표를 기준으로 25×25 크기로 얻은 이미지 패치는 좌표와 함께 Vision Transformer[9,10] 네트워크의 입력으로 활용된다. Transformer 네트워크를 선정한 이유는 일반적인 합성곱 신경망들과는 다르게 입력 정보의 차원 축소를 하지 않아서 정보 손실이 적고, 시계열 측면이나 공계열 측면 모두를 각자의 영역에 맞춰서 포괄적으로 학습이 가능하기 때문이다. 정보의 손실 없이 모두 학습하므로 전체 이미지를 사용하기보다는 지역적인 정보를 얻어야 하는 관절 좌표 이미지 패치 스트림에 적절하다. 네트워크의 전반적인 구조는 그림 7과 같다. 입력으로 좌표와 패치를 같이 받게 되는데, 먼저 패치들만 따로 Spatial Vision Transformer를 진행한다. 이후에 Temporal Vision Transformer를 진행하기 전에 좌표들을 임베딩한 벡터 값을 패치 벡터 앞에 연결해서 학습을 진행한다. 다른 스트림들과 마찬가지로 정확도 점수를 마지막 앙상블 층으로 보내어 최종 결과를 얻는다.



(a) T-1 (b) T (c) T′

그림 6. 좌표 채우기 예시 (T: 시간)
Fig. 6. Example of joint point merging (T: Time)

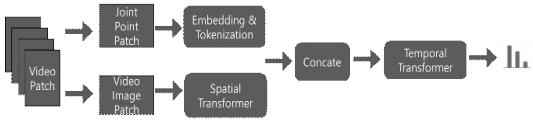


그림 7. 관절 좌표 이미지 패치 트랜스포머의 구조
Fig. 7. Structure of joint point image patches transformer

2.4 스켈레톤 스트림

수화자의 스켈레톤 정보를 얻기 위한 스트림이다. 유사한 수화에서 손의 위치나 손의 이동 정도 등에 따라 뜻이 달라지는 경우가 있기 때문에 수화자의 손, 손가락, 얼굴 등의 위치에 대한 정보가 중요하다. 공간적인 부분에서는 손과 손가락이 어디에 있는지, 시간적인 부분에서는 손이 얼마나 움직이는지 알기 위해 행동 인식에서 좋은 성능을 보인 Spatial Temporal 그래프 합성곱 신경망 (STGCN)[11]을 활용한다. 그래프를 만들기 위해 앞서 얻은, 얼굴에 7개, 한 손에 21개씩, 총 49개의 좌표를 사용한다. 얼굴, 양손 순으로 연결하여 식 (1)과 같은 형태로 하나의 벡터를 구성한다.

이렇게 얻은 벡터와 인접 행렬을 그림 8의 입력으로 활용해서 공간적인 연관성에 대한 정보를 얻고, 시간에 관해 학습하여 수화를 구분한다. 다른 두 개의 스트림과 마찬가지로 정확도 점수를 마지막 앙상블 층으로 보내어 인식 과정을 마무리한다.

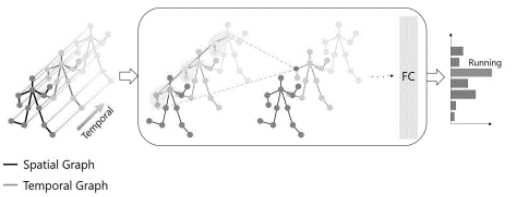


그림 8. ST-GCN [11]
Fig. 8. ST-GCN

Ⅲ. 실 험

3.1 실험 세부사항

본 논문에서 사용한 데이터 세트는 단어 단위의 수화 인식 데이터 세트 중에서 가장 크고 많은 수화를 포함하고 있는 WLASL[1]이다. 이 데이터 세트에는 총 4가지의 세트가 존재하는데, 클래스가 각각 100개, 300개, 1,000개, 2,000개가 있고, 각각 WLASL100, WLASL300, WLASL1000, WLASL2000이라고 한다. 포함된 동영상의 개수는 표 1과 같다. 본 논문에서는 이 중에서 WLASL100만 사용하였다. 각 영상은 32개의 프레임으로 구성되어 있다. 해당 데이터 세트에 포함된 수화 영상들은 수화자가 영상 가운데에 위치해 있다. 또한, 주변에는 수화 인식에 필요 없지만 영상의 다양성을 위해서 서로 다른 배경이 따로 존재한다.

제안하는 방법의 학습은 각 스트림 별로 다르게 하였다. 기본 스트림은 입력 데이터로 일반 RGB image를 활용하였고, 학습 방법도 [1]과 같은 방식을 사용하였다. 기본 스트림에서는 이미지들을 256×256의 일정한 크기로 변경하고, 학습 시에는 224×224의 크기로 무작위로 잘라서 사용한다. 평가 시에는 중심을 기준으로 224×224의 크기로 잘라서 사용한다. 관절 좌표 이미지 패치 스트림은 동영상의 한 프레임당 49개의 좌표를 기준으로 각각 25×25 크기의 패치를 하나의 프레임 입력으로 사용한다. 마지막 스켈레톤 스트림은 관절 좌표로 만든 그래프와 이에 대한 인접 행렬을 입력으로 활용한다. 모든 스트림에서 Adam을 활용하고, 손실함수로는 Cross Entropy를 사용하였다. 학습 데이터와 실험 데이터는 8대 2 비율로 나누어 Cross-Validation을 수행하였으며, 학습은 총 200회 진행하였다.

표 1. WLASL 데이터 세트에 관한 정보
Table 1. Information of WLASL dataset

Subset	#class	#Video	#Signer
WLASL100	**100**	**2,038**	**97**
WLASL300	300	5,117	109
WLASL1000	1,000	13,168	116
WLASL2000	2,000	21,038	119

3.2 패치 크기에 따른 결과 비교

관절 좌표 이미지 패치는 관절 좌표를 중심으로 이미지 패치를 만든다. 이 때 고려해야 할 것 중 하나가 패치의 크기이다. 총 3 가지의 경우로 실험하였다. 트랜스포머에서 가장 자주 사용하는 크기인 16, 그의 배수인 32, 그리고 그 중간 값 중 하나인 25를 선정해서 비교하였다. 자세한 결과는 표 2에서 확인 가능하며, 패치의 크기에 따라 P16, P25, P32라고 명명하였다. 데이터는 WLASL100을 사용하였고, 관절 좌표의 개수는 49로 설정하였으며, 모든 네트워크 설정은 동일하게 실험을 진행하였다. 결과적으로 패치 크기가 25인 경우에서 가장 좋은 성능인 70.28%가 나왔다. 패치 크기 16인 경우는 64.73%로 어느 정도 좋은 성능을 보였지만, 패치 크기가 32인 경우에는 오히려 성능이 저하되는 모습을 보였다. 이 결과로 알 수 있는 점은 패치의 크기에 의해 패치끼리 어느 정도 겹치는 부분이 생기는 것은 성능을 향상시키만, 패치의 크기가 너무 작아서 안 겹치거나, 반대로 너무 커서 많이 겹치는 경우에는 성능이 떨어지는 것을 확인할 수 있다.

표 2. 패치 크기에 따른 결과 비교
Table 2. Result comparison according to patch size

#Patch Size	Accuracies (%)
P16	64.73
P25	**70.28**
P32	60.18

3.3 관절 좌표 개수에 따른 결과 비교

처음 관절 좌표를 얻으면 총 55개의 좌표가 얻어진다. 표 3을 보면 좌표의 개수에 따라 결과가 다른 것을 확인할 수 있다. N29는 얼굴에 7개의 좌표, 한 손에 11개 좌표로 총 29개의 좌표를 사용한 것이며, N49는 얼굴 좌표 7개와 한 손에 21개씩, 총 49개의 자료를 사용한 것이다. N55는 얼굴에 13개, 한 손에 21개씩, 총 55개의 좌표를 사용한 것이다. N29와 N49, N55의 가장 큰 차이점은 얼굴 좌표 13개 중에서 몇 개의 좌표를 포함하는가이다. N55는 모든 좌표를, N49와 N29는 중복을 최소화한 7개의 좌표를 선정하였다. 데이터 세트는 WLASL100을 활용하고 패치 크기는 3.2절에서 확인된, 성능이 가장 좋은 25로 설정하였으며, 네트워크는 관절 좌표 이미지 패치 스

표 3. 관절 좌표 개수에 따른 결과 비교
Table 3. Result comparison according to number of joint points

#Joint Point	Accuracies (%)
N29	64.7
N49	**70.28**
N55	63.1

트림의 네트워크를 사용하였다. N29는 최종 결과에서 64.7%가 나왔고, N55는 63.1%가 나왔다. 마지막으로 N49는 70.28%가 나와서 최종적인 다중 네트워크에는 N49를 사용한다. 이러한 결과를 통해서 알 수 있는 점은 수화 인식에서 중요한 손의 정보는 오히려 많은 겹침이 성능에 좋은 영향을 주지만, 중요도가 비교적 떨어지는 얼굴의 눈, 귀는 겹침이 적지만 중요 좌표는 포함된 경우가 성능에 더 좋은 영향을 준다는 것이다.

3.4 기존 결과와의 성능 비교

기존 논문의 결과[1,3,7]과 비교하여 표 4에 자세하게 표시하였다. 다중 스트림 네트워크 뿐만 아니라, 관절 좌표 이미지 패치의 성능을 검증하기 위해 기존 논문의 단일 스트림 네트워크도 성능을 비교하였다. 이미지 기반 방법론의 기본이 되는 I3D[1]와 자세 기반 방법론의 기본이 되는 Pose-TGCN[1]과 비교해보면 약 17%와 27% 성능이 향상되었다. 다중 네트워크 방법[3]과 비교하면 기존 방법의 I3D+ST-GCN(5)는 RGB Image와 Local Image를 활용한 스트림 4개와 스켈레톤 정보를 이용한 스트림 1개, 총 5개의 스트림을 사용하여 약 77.5%의 정확도를 얻었다. 이에 flow 데이터를 추가해서 총 6개의 스트림을 사용한 I3D+ST-GCN(6)은 약 81%의 정확도를 얻었다. 제안하는 방법은 3개의 스트림을 사용하면서도 기존 방법의 I3D+ST-GCN(5)에 비해 약 5%의 성능이 향상되었고, I3D+ST-GCN(6)에 비해서는 약 1%정도의 성능이 향상되었다.

표 4. 다중 스트림 네트워크 성능 비교
Table 4. Result comparison of multi-stream networks

Reference	Network(#stream)	Input Data	Acc
[1]	I3D (1)	Image	65.89
	Pose TGCN (1)	Skeleton	55.43
[7]	Transformer (1)	Skeleton	63.18
[3]	I3D (4)	Image + Local	76.60
	I3D + ST-GCN (5)	Image + Local + Skeleton	77.48
	I3D + ST-GCN (6)	Image + Flow + Local + Skeleton	81.38
Ours	Transformer (1)	Joint Point Patch	70.28
	Transformer + I3D (2)	Joint Point Patch + Image	78.8
	Transformer + I3D + ST-GCN (3)	**Joint Point Patch + Image + Skeleton**	**82.3**

Ⅳ. 결 론

본 논문에서는 수화 인식에 필요한 지역적인 정보를 더욱 효과적으로 얻기 위한 관절 좌표 이미지 패치와 패치의 연관성을 학습하기 위한 다중 네트워크에 대한 연구를 진행하였다. 추후 연구로는 기본 스트림을 더 좋은 성능을 보이는 네트워크로 변경하고, 세부적인 관절 좌표를 선정하는 방법에 대한 연구가 필요하다. 또한, 스켈레톤 스트림 자체의 성능을 높이는 방법에 대한 연구도 필요하다.

References

[1] 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 *Proc. 2020 IEEE WACV*, pp. 1448-1458, Snowmass Village, CO, USA, Mar. 2020. (https://doi.org/10.48550/arXiv.1910.11006)

[2] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, “Learning to zoom: A saliency-based sampling layer for neural networks,” in *Proc. ECCV*, pp. 51-66, Munich, Germany, Sep. 2018. (https://doi.org/10.48550/arXiv.1809.03355)

[3] M. Maruyama, S. Ghose, K. Inoue, P. P. Roy, M. Iwamura, and M. Yoshioka, “Word level sign language recognition with multi-stream neural networks focusing on local regions,” *arXiv:2106.15989*, 2021. (https://doi.org/10.48550/arXiv.2106.15989)

[4] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proc. 2017 IEEE Conf. CVPR*, pp. 4724-4733, Honolulu, HI, USA, Jul. 2017. (https://doi.org/10.48550/arXiv.1705.07750)

[5] Y. Du, P. Xie, M. Wang, X. Hu, Z. Zhao, and J. Liu, “Full transformer network with masking future for word-level sign language recognition,” *Neurocomputing*, vol. 500, pp. 115-123, Aug. 2022. (https://doi.org/10.1016/j.neucom.2022.05.051)

[6] A. Tunga, S. V. Nuthalapati, and J. Wachs,

"Pose-based sign language recognition using GCN and BERT," in *Proc. 2021 IEEE WACVW*, pp. 31-40, Waikola, HI, USA, Jan. 2021. (https://doi.org/10.48550/arXiv.2012.00781)

[7] M. Bohácek and M. Hrúz, "Sign pose-based transformer for word-level sign language recognition," in *Proc. 2022 IEEE/CVF WACVW*, pp. 182-191, Waikola, HI, USA, Jan. 2022. (https://doi.org/10.1109/WACVW54805.2022.00024)

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211-252, Dec. 2015. (https://doi.org/10.48550/arXiv.1409.0575)

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViVit: A video vision transformer," *arXiv:2103.15691*, 2021. (https://doi.org/10.48550/arXiv.2103.15691)

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th ICLR*, May 2021. (https://doi.org/10.48550/arXiv.2010.11929)

[11] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artificial Intell.*, pp. 7444-7452, New Orleans, LA, USA, Feb. 2018. (https://doi.org/10.48550/arXiv.1801.07455)

강 현 석 (Hyeon-Seok Kang)

2021년 2월 : 광운대학교 로봇학부 학사

2023년 2월 : 광운대학교 로봇학과 석사

<관심분야> 수화 인식, 머신러닝

[ORCID:0009-0006-3131-0425]

박 광 현 (Kwang-Hyun Park)

1994년 2월 : KAIST 전기및전자공학과 학사

1997년 2월 : KAIST 전기및전자공학과 석사

2001년 2월 : KAIST 전기및전자공학과 박사

2008년 3월~현재 : 광운대학교 로봇학부 교수

<관심분야> 기계학습, 로봇 소프트웨어

[ORCID:0000-0003-3041-4055]