공 학 석 사 학 위 논 문

트랜스포머와 샴 네트워크를 이용한 EOG 기반 눈 글씨 인식

2023년 8월

부 경 대 학 교 대 학 원

인 공 지 능 융 합 학 과

강 동 현

공 학 석 사 학 위 논 문

트랜스포머와 샴 네트워크를 이용한 EOG 기반 눈 글씨 인식

지도교수 장 원 두

이 논문을 공학석사 학위논문으로 제출함.

2023년 8월

부 경 대 학 교 대 학 원

인 공 지 능 융 합 학 과

강 동 현

강동현의 공학석사 학위논문을 인준함.

2023년 8월 18일

위 원 장 공학박사 한 영 선 (인)

위 원 공학박사 김 훈 희 (인)

위 원 공학박사 장 원 두 (인)

목 차

표 목 차

그 림 목 차

Recgnition of EOG-based Eye-written Characters using Transformers and Siamese Network

Dong Hyun Kang

Department of Artificial Intelligence Convergence.
The Graduate School, Pukyong National University

Abstract

As bio-signal measurement technology develops, interest in bio-signal-based HCI (Human-Computer Interface) is increasing. Among them, applications based on electrooculogram (EOG) are used as a means of communication for patients with degenerative neurological syndromes or quadriplegia. Various studies have been conducted on recognizing eye-written characters using EOG. However, EOG-based handwriting data is limited by the small sample size of available data due to collection constraints.

In this study, we paid attention to the limited number of data. To solve this problem, we proposed a recognition model that combines the idea of Reference data and VIT (vision transformers) based Siamese networks. The Siamese network is a learning methodology that determines the class matching of two inputs. we apply ViT which is suitable for time-series data analysis. We introduce Reference data to transform Siamese networks, which deal with binary classification tasks that only determine whether input pairs match classes, into unambiguous class prediction and multiple classification tasks.

In this study, we use 10 Arabic numeral datasets and 12 Katakana stroke datasets. The experimental results show that the ViT-based Siamese network achieves high performance, with recognition accuracy of up to 91.9% on the Arabic numerals dataset and up to 84.7% on the Katakana strokes dataset. We found that the Siamese network has robust recognition performance, reaching about 90% accuracy, except for some classes in zero-shot learning, which is one of the main features of Siamese networks.

1. 서 론

최근 생체신호 측정 기술의 발달로 이를 활용한 인간-컴퓨터 인터페이스(HCI, human-computer interface)를 통해 실시간 상호작용을 위한 시도가 늘고 있다. 생체 신호는 신체 활동에 따라 발생하는 미세전류로서 근전도(EMG, electromyogram), 뇌전도(EEG, electroencephalogram), 안구전도(EOG, electrooculogram) 등이 있다[1]. 그 중 안구전도 기반 인간-컴퓨터 인터페이스는 퇴행성 신경 증후군 혹은 사지마비 환자들의 의사소통 수단이다. 이는 신체 마비 과정에서 안구운동에 필요한 근육 파괴를 야기하지 않기 때문에 가능한 일이다[2].

안구전도는 각막과 망막에서 발생하는 전위를 측정하는 방법이며, 전극의 부착 위치에 따라 다양한 표현이 가능하다[3]. 안구전도 기반 시선 추적 시스템은 카메라를 활용한 방법에 비해 주변 환경이나 신체적 상태 등의 조건에서 상대적으로 자유로우며, 비교적 저렴한 비용과 쉬운 사용성이 장점이다[4]. 이러한 이유로 최근 수년간 안구전도 기반 인간-컴퓨터 인터페이스를 활용한 안구 움직임 추적 기법, 시스템 등의 다양한 연구 결과와 응용 프로그램이 제안되고 있다[5, 6. 7]. 하지만, 안구전도를 활용한 눈 깜빡임 인식[8], 안구 움직임 인식[5] 등의 어플리케이션은 한정된 표현만을 인식하기 때문에 의사 전달의 다양성은 여전히 부족하다.

이러한 이유로 안구전도를 활용한 눈 글씨(eye writing characters) 인식에대한 연구가 주목받고 있다. 눈 글씨는 사용자의 안구 움직임에 따라 나타내는 시선의 위치 변화에 따른 경로를 글로써 활용하는 것이다. Tsai 등은 휴리스틱(heuristic) 기반의 인식 시스템을 제안하였으며, 10종의 아라비아 숫자와 4가지의 연산자로 이루어진 문자를 75.5%의 정확도(believability)로 인식하였다[9]. Lee 등은 26종류의 알파벳과 3개의 기능(space, backspace, enter) 문자를 인식하기 위해 DTW(dynamic time warping)를 적용하였으며 87.38의 F1-score를 달성하였다[10]. Chang 등은

10종의 아라비아 숫자로 구성된 문자들을 DPW(dynamic positional warping)와 서포트 벡터 머신(SVM, support vector machine)을 결합한 방법을 사용하여 94.37%의 인식 정확도를 달성하였다[11].

최근 주목받는 인공지능 모델인 딥러닝 기술을 적용한 사례도 있다. Fang 등은 DNN(deep neural network)과 HMM(hidden markov model)을 결합한 방법으로 12종류의 가타카나 문자 인식을 시도해 86.5% 인식률을 달성하였다[12]. Chang 등은 인셉션(Inception) 구조[13]를 기반으로 둔 딥러닝 모델을 제안하였으며, 10종의 아라비아 숫자 표현들을 97.78%의 정확도로 분류하였다[14].

안구전도 기반 눈 글씨 인식을 목적으로 제안된 기존 연구들의 공통적인 문제점은 활용 가능한 데이터의 표본이 적다는 것이다. 이는 몇 가지 원인이 존재한다. 가장 주요한 원인은 신호를 수집할 때 잡음(noise)과 인공물(artifact)이 유입된다는 것이며, 이는 정교한 데이터의 수집에 어려움을 야기한다. 구체적으로 안구전도는 신호원에 부착된 전극을 사용해 수집되며, 이때 다른 기관으로부터 유입되는 잡음과 인공물은 신호의 안정성에 부정적인 영향을 준다. 유입된 잡음은 고속 푸리에 변환, 웨이블렛 변환 등 전통적인 방법으로 제거할 수 있지만, 그에 따른 전처리 시간이 소모된다. 또한, 수집되는 잡음이 넓은 주파수 대역에 분포하는 경우 완벽한 제거가 어렵다[7]. 또 다른 원인은 안구전도 수집 시 피험자의 피로도나 심리 상태, 주변 환경에 민감하게 반응한다는 것이며, 이는 신호의 변화에 큰 영향을 준다. 또한, 눈 글씨 수집 과정은 눈 깜빡임 혹은 단순한 움직임에 비해 상대적으로 긴 시간을 소모하기 때문에 해당 조건들에 더 취약하다. 이와 같은 기술적 문제와 더불어 개인정보보호 등과 관련하여 법률적·도덕적 규제 또한 데이터 수집에 제한을 주는 한 요인으로 작용한다.

활용 가능한 데이터가 부족하다는 문제에 주목하여 Kang 등은 샴 네트워크(siamese network) 기반의 딥러닝 모델을 제안하였다[15]. 해당 연구는 샴 네트워크 이론[16]을 활용해 다중 클래스 분류 문제를 이진 분류 문제로 재정의함으로써 소량의 데이터만을 사용하면서도 높은 인식률을 달성할

수 있다는 것을 증명하였다. 해당 연구에선 10종류의 아라비아 숫자 데이터를 540개 사용하였으며, 학습 데이터와 평가 데이터를 2:1의 비율로 나누어 학습한 실험에서 94.44%의 분류 정확도를 달성했다. 이와 동시에 학습 데이터와 평가 데이터를 1:10의 비율로 나눈 실험에서도 준수한 성능(80.80%)을 보였다. 하지만, 눈 깜빡임(eye blink)이 제거된 신호만을 사용했다는 점에서 샴 네트워크가 실용적인 방법인지에 대한 객관성을 보장하기엔 부족함이 있었다. 또한, 샴 네트워크의 특징 중 하나인 학습되지 않은 클래스 인식(zero-shot learning)[16]은 해당 연구에서 유의미한 효과를 보이지 않았으며, 이는 눈 글씨 클래스 수가 증가함에 따라 주기적인 재학습이 필요하다는 한계를 제시하였다.

본 연구에선 샴 네트워크를 활용한 눈 글씨 인식의 실용적인 측면과 한계에 주목하고, 기존의 연구[15]를 개선해 문제를 해결하고자 한다. 구체적으로 학습 데이터의 클래스를 추가함으로써 데이터의 표본을 증강해 학습되지 않은 클래스의 인식 정확도 상승을 도모한다. 또한, 눈 깜빡임이 제거된 신호만을 사용한 기존의 연구와는 달리 최소한의 전처리가 수행된 신호를 사용함으로써 실용성 측면을 보완한다. 이와 동시에 샴 네트워크 내 특징 추출기의 성능을 개선하고자 한다. 본 연구는 결과적으로 샴 네트워크를 활용함으로써 전처리 과정을 최소화할 수 있다는 점에서 학술적 의의가 있다. 이와 더불어 실무적 측면에선 시간적·물리적 비용 절감의 가능성을 제시하며 생산성 향상에 기여할 수 있음을 기대한다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 눈 글씨 데이터의 특징 및 전처리 과정, 주요 알고리즘을 살펴본다. 제 3장에서는 세부적인 실험 방법 설명과 결과를 분석한다. 마지막 제 4장에선 연구의 종합적인 결론과 한계 및 향후 연구 방향, 시사점 등을 제시한다.

2. 연구 방법

제 2장에서는 본 연구의 연구 방법이 소개된다. 본 장은 크게 연구에 사용된 데이터 셋과 전처리 기법, 주요 알고리즘인 샴 네트워크에 대한 설명으로 구성된다. 연구에서 제안하는 방법론의 전체적인 구조도는 Fig.1과 같다. 구체적으로 우선 본 연구에서 사용된 두 종류의 눈 글씨 데이터 셋에 대한 설명과 딥러닝 입력에 적합한 전처리 과정을 다룬다. 다음, 샴 네트워크에 대한 설명과 입력 데이터의 특징을 추출하기 위한 딥러닝 모델, 학습 방법과 최적화 방안을 기술한다.

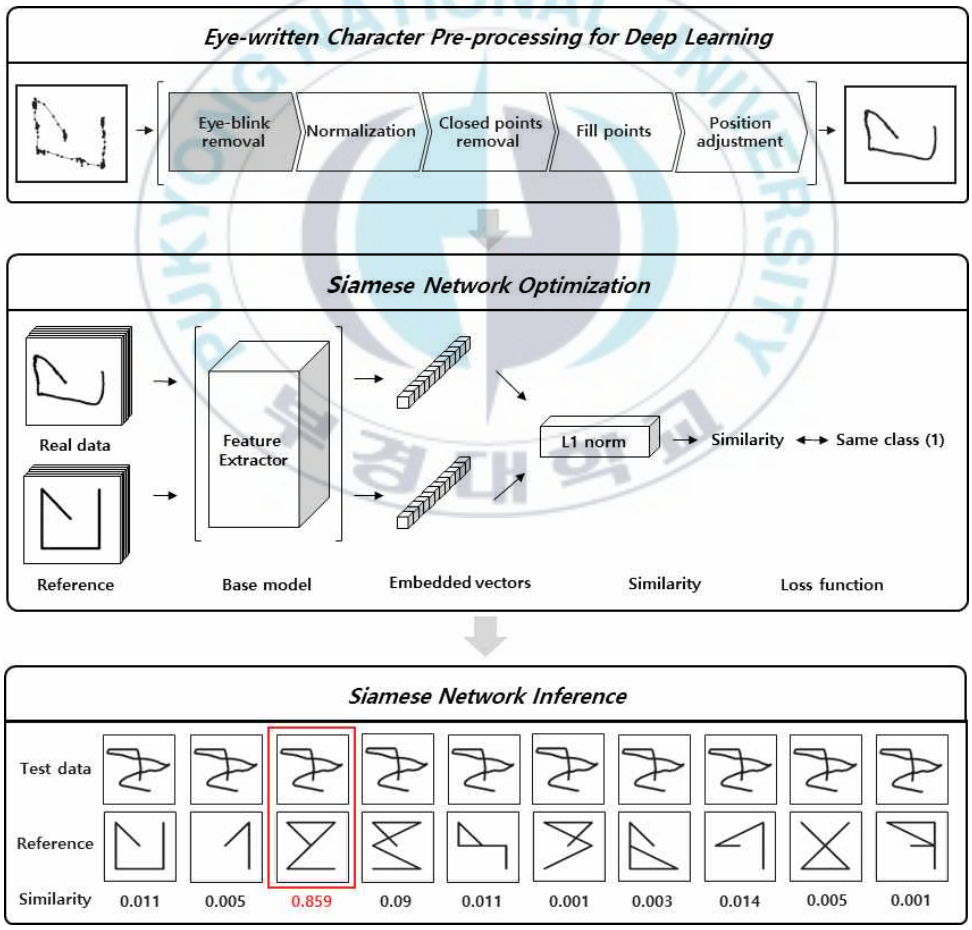

Fig 1. Overall of proposed mechanism

2.1 데이터 셋

본 연구에선 공개 데이터인 10종류의 아라비아 숫자[11]와 12종류의 가타카나 문자[12] 데이터를 사용하며, 샴 네트워크 적용을 위한 참조 데이터(Reference data)를 사용한다.

아라비아 숫자 데이터 셋은 18명의 참가자로부터 0-9까지의 문자를 수집한 것이다(Fig. 2). 참가자들은 각 문자를 3회 반복 시행하였으며, 데이터는 총 540개의 신호로 구성되어있다. 본 연구에서는 수집된 원신호에서 몇 단계 처리 과정을 거친 후의 신호를 사용한다. 2048Hz 주기로 수집된 원신호는 64Hz로 변환(down sampling) 후 미디언 필터(median filter)를 적용하여 잡음이 제거되었으며, 선형회귀를 사용해 저대역 주파수에 존재하는 드리프트가 제거되었다. 기존의 연구[15]에선 잡음, 드리프트, 눈 깜빡임이 제거된 신호를 사용하였으나, 본 연구에선 눈 깜빡임 제거의 불필요성을 증명하기 위해 눈 깜빡임 제거 전·후의 신호에 대한 실험을 진행한 후, 이후의 실험에선 더 효율적이라 판단되는 신호를 사용한다.

가타카나 문자 데이터 셋은 12개의 획을 표현하였으며, 6명의 참가자로부터 수집되었다(Fig. 3). 참가자들은 각 문자를 10회 반복 입력하였으며, 총 720개의 신호로 이루어져 있다. 본 연구에선 원신호에 FIR 저역통과 필터(finite impulse response low pass filter)이 적용되어 주기가 20Hz 이하인 잡음 영역이 제거되고, DC blocker이 적용되어 드리프트까지 제거된 후의 데이터를 사용한다.

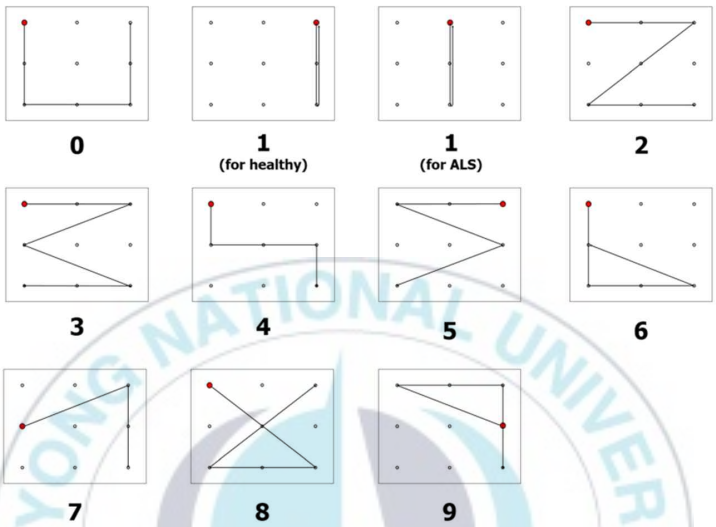

Fig 2. Eye written patterns of Arabic numbers [11]

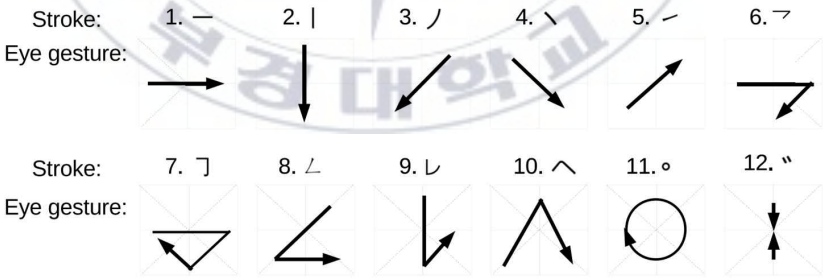

Fig 3. Eye written patterns of Katakana gestures [12]

샴 네트워크는 두 입력이 같은 클래스인지 아닌지를 판단하는 이진 분류 알고리즘이다. 이는 입력 간 유사도를 계산할 뿐 클래스의 특징을 학습하는 것은 아니기 때문에 다중 분류 문제에 적합한 기법은 아니다. 본 연구에선 각 클래스에 대한 명확한 기준을 제시하기 위해 참조 데이터(reference data)를 추가한다(Fig. 4, 5). 이는 이진 분류 문제에 강한 성능을 보이는 샴 네트워크를 다중 분류 문제로 접근할 수 있도록 한다. 또한, 참조 데이터를 학습 데이터에 포함함으로써 데이터 증강과 함께 네트워크가 눈 글씨 데이터에 과적합 되는 것을 방지하는 효과가 있다.

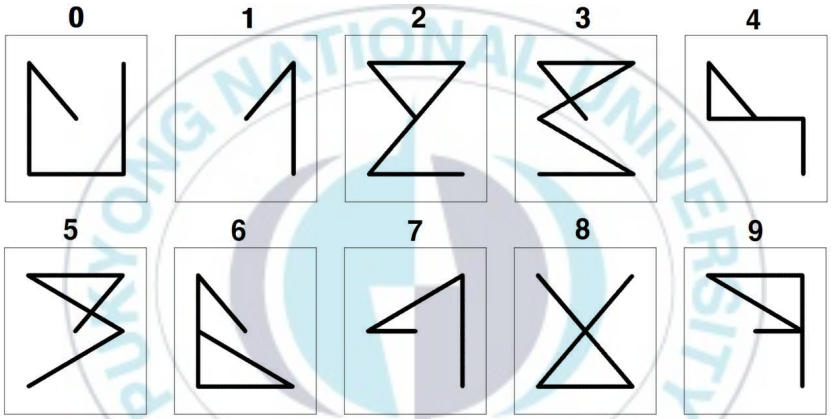

Fig 4. Reference data of Arabic numbers

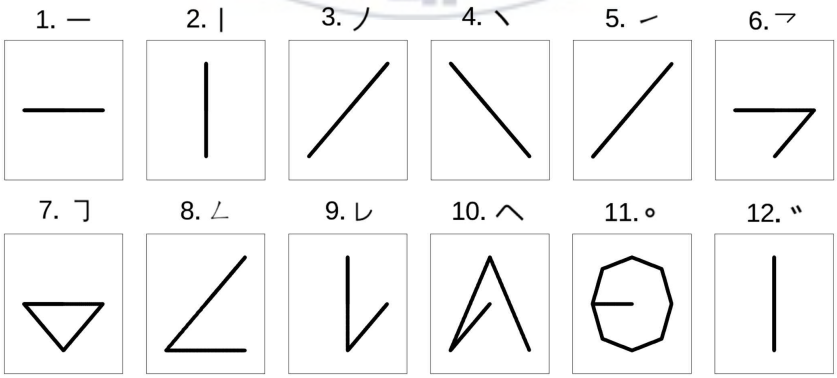

Fig 5. Reference data of Katakana gestures

2.2 전처리

잡음과 드리프트가 제거된 신호는 딥러닝 모델에 적합하도록 전처리 과정을 거친다. 일반적인 시계열 데이터는 시간에 대한 특징이 중요한 요소로서 작용한다. 그러나 눈 글씨의 경우 시계열 데이터이지만, 2차원 공간에서의 공간적 특성이 중요하기 때문에 데이터 내의 순서 정보를 제외한 시간적 특징은 무시하도록 한다. 이러한 이유로 본 연구에서 전처리는 눈 깜빡임 제거, 정규화(normalization), 데이터 재배치(data reconstruction) 순으로 이루어진다. 눈 깜빡임 제거는 MSDW (maximum summation of the first derivative within a window)[17] 기법을 적용하며, 본 연구에선 실험에 따라 해당 과정을 수행한 신호와 수행하지 않은 신호를 사용한다.

눈 깜빡임 제거 과정 후, 신호는 정규화를 거친다. 일반적인 정규화는 0과 1 사이의 값으로 변환되어 시작점이 (0.5,0.5)로 되지만, 본 연구에선 시작점이 (0,0)이 되도록 -1과 1 사이의 값으로 정규화한다. 이는 2차원 데이터임을 고려하여 일반적인 정규화에 비해 4배 넓은 범위에서 위치적 특징을 추출할 수 있다는 장점이 있다.

정규화된 신호는 3가지 단계로 구성된 데이터 재배치 과정을 거친다. 첫 번째 단계는 중첩 좌표 제거(closed points removal)이다. 이는 데이터 수집 시 고정된 영역에서 중첩 입력되어 발생한 불필요한 좌표들을 제거하는 과정이다. 먼저, 일정한 주기로 수집된 연속된 좌표를 동일한 크기로 그룹화한다. 그룹화된 좌표 집합을 순회하며 좌표 간 거리를 측정해 중첩 영역 여부를 판별한다. 중첩 영역이라 판단되는 구간의 좌표들을 해당 구간의 중심 좌표로 치환한다. 두 번째 단계는 좌표 추가 과정(fill points)이다. 해당 과정은 좌표 간 간격을 줄이는 것이 목적이다. 예를 들어, Fig. 7에서 중첩 좌표 제거가 이루어진 신호의 경우 문자의 끝단(edge)에 위치한 좌표들에 비해 간선(line)에 위치한 좌표들의 수가 적은 것을 확인할 수 있으며, 이를 교정해야 한다. 해당 단계에선 목표 좌표 수(신호의 길이)의 임의

의 N배에 달하는 좌표를 추가한 후, 반복적으로 반감한다. 마지막 단계는 위치 조정(position adjustment)이다. 이는 좌표 추가 과정이 수행되었음에도 좌표 간 간격에 미세한 차이를 제거하기 위해 수행된다. 구체적으로 순서가 있는 좌표 사이에서 두 좌표 간 간격을 일정하게 조정한다.

데이터 재배치는 기존의 연구에서 수행된 5단계의 데이터 리샘플링(data resampling) 과정이 3단계로 간소화된 과정으로서 이에 소모되는 전처리 시간을 절약할 수 있다(Fig. 6, 7).

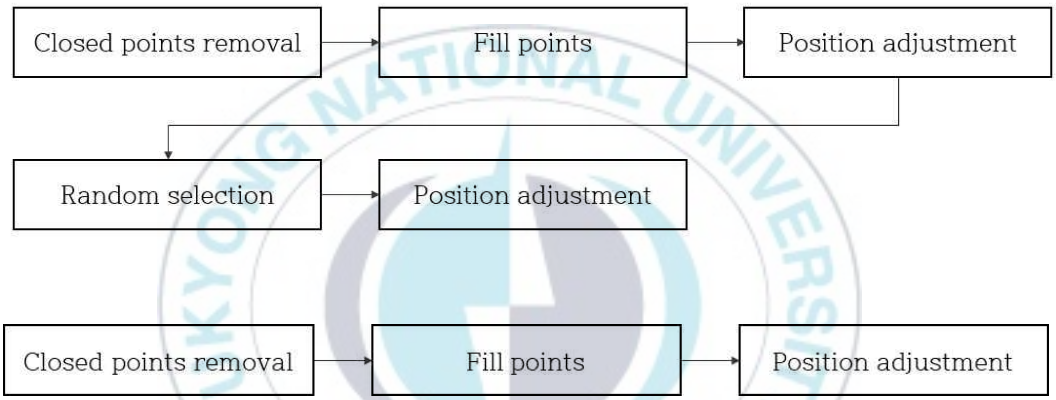

Fig 6. Previous data resampling(top) process and data reconstruction process(bottom) that streamlined

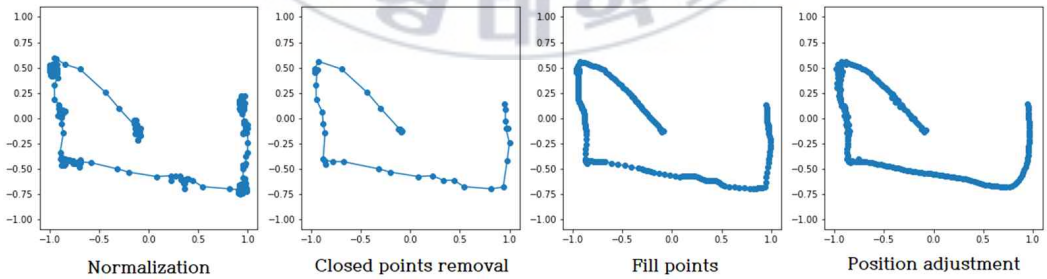

Fig 7. Illustration of data reconstruction with class 0 of Arabic numbers

2.3 샴 네트워크

2.3.1 네트워크 구조

본 연구의 주요 알고리즘은 샴 네트워크이다. 샴 네트워크는 학습해야 할 클래스 수에 비해 개별 클래스의 표본이 현저히 부족할 경우 적합한 해결 방법이 될 수 있다[16]. 이는 두 입력 데이터의 유사도를 계산하여 같은 클래스인지 아닌지는 판단한다[18-19]. 예를 들어, 본 연구에서 사용하는 아라비아 숫자 데이터는 Fig. 8과 같이 쌍을 이루어 표현할 수 있다.

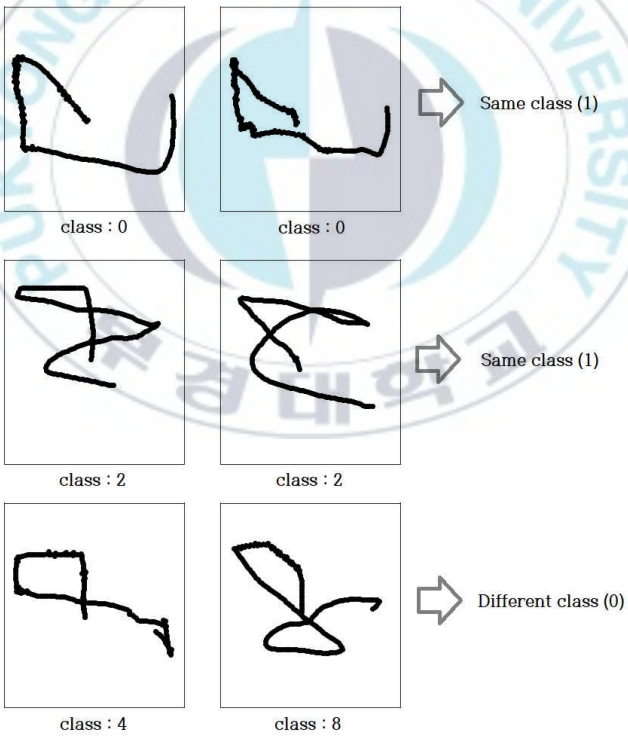

Fig 8. Example pairs of eye writing characters for Siamese network

샴 네트워크의 구조는 Fig. 9과 같다. 한 쌍으로 입력된 두 데이터는 샴 네트워크 내 특징 추출기를 통과 후 특징 벡터로 변환된다. 두 입력에 대한 특징 벡터 쌍은 L1 norm 혹은 L2 norm 등 거리 함수가 적용되어 유사도 평가가 이루어지며, 본 연구에선 L1 norm을 적용한다. 유사도는 최종적으로 시그모이드 함수(sigmoid function)가 적용되어 0과 1 사이의 값으로 변경되어 출력된다. 이후 샴 네트워크는 출력된 유사도를 기준으로 두 입력의 클래스가 같다면 거리가 최소가 되도록, 다르다면 최대가 되도록 최적화된다.

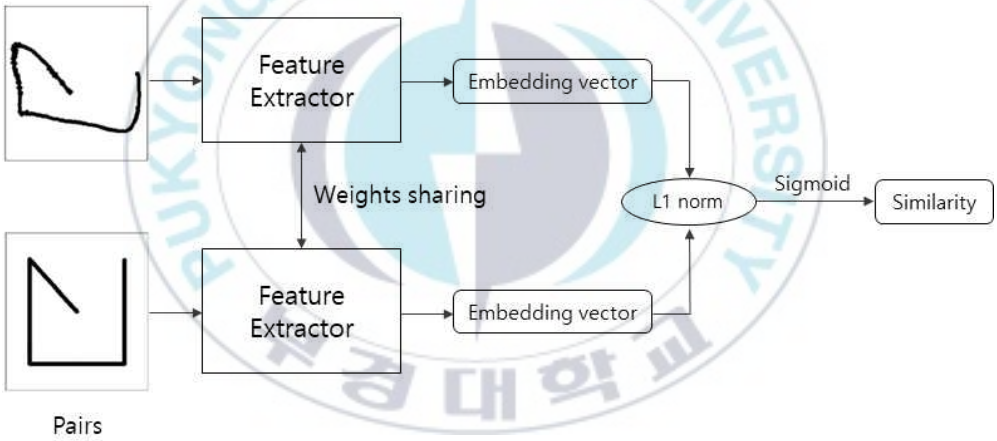

Fig 9. Architecture of Siamese network

2.3.2 특징 추출기

본 연구에서 사용하는 특징 추출기는 ViT(vision transformers)와 1D CNN+LSTM 모델이다.

본 연구에서 제안하는 ViT 모델은 이미지 인식 분야에서 널리 사용되는 ViT 모델을 시계열 데이터에 맞게 변형한 구조를 가진다. 이미지 인식에 활용되는 ViT는 자연어 처리 분야에서 제안된 트랜스포머의 인코더 구조만을 활용한 변형한 모델이다. 구체적으로 2차원 형태인 이미지를 패치(patch)들로 나눈 후, 패치 별로 순번을 부여하여 이미지의 특징 맵(feature map)을 추출한다. 예를 들어, 128×128 이미지는 32×32 크기의 패치 16개로 분할되어 순번이 부여된 후, N개의 트랜스포머 인코더에 입력되어 특징을 학습한다[20]. 본 연구에 제안하는 ViT 특징 추출기는 1차원 형태인 시계열 데이터를 입력받는다. 이때 일반적인 트랜스포머와 차이점은 입력되는 모든 데이터가 아닌 패치 단위로 분할되어 연산이 수행된다는 것이다. 구체적으로 연구에서 사용한 ViT 특징 추출기는 길이가 200인 신호를 각각 연산하는 것이 아닌 10개의 패치로 분할 한 후, 순번을 부여해 특징 벡터를 추출한다(Fig. 10).

CNN+LSTM 특징 추출기는 저자의 이전 연구[15]에서 사용한 네트워크 구조와 동일하다. 구체적으로 2개의 1D CNN 층과 2개의 LSTM 층, 그리고 계층적 어텐션 매커니즘 층으로 구성되어있다(Fig. 11). ViT 특징 추출기는 CNN+LSTM 특징 추출기와 성능 비교가 진행된다.

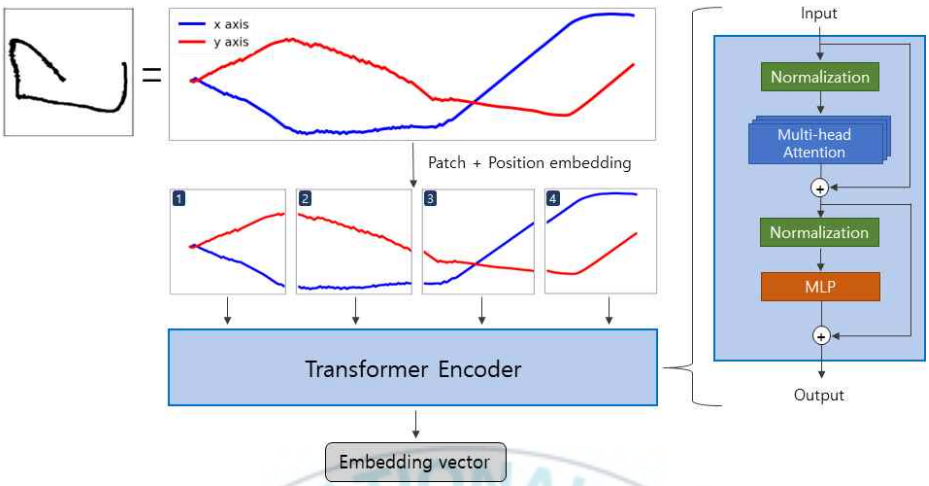

Fig 10. Custom ViT(vision transformers) architecture for eye writing characters when patch size is 4

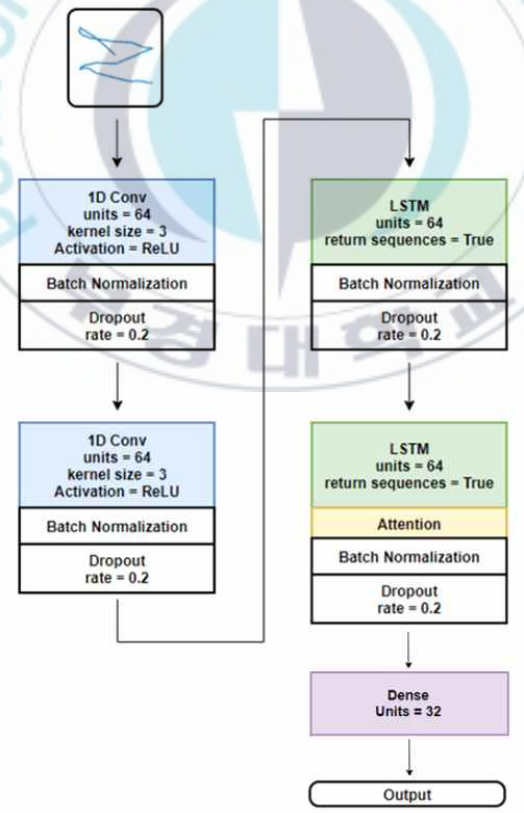

Fig 11. CNN+LSTM based feature extractor[15]

2.3.3 네트워크 입력

샴 네트워크는 여러 데이터 쌍으로 구성된 배치(batch)를 입력받는다. 네트워크 학습을 위해 입력되는 학습 배치는 지정된 배치 크기(batch size)만큼의 쌍으로 이루어진다. 하나의 배치에는 서로 같은 클래스의 쌍(pair)들과 다른 클래스의 쌍들이 동일한 비율로 구성된다. 본 연구에선 참조 데이터의 포함 여부에 따라 두 종류의 학습 배치와 하나의 평가 배치를 사용한다. 기본 학습 배치는 눈 글씨 데이터로만 구성된 배치로써 네트워크는 수집된 눈 글씨 데이터에 대한 특징을 학습한다. 또 다른 학습 배치는 참조 데이터가 포함되며, 이는 네트워크가 눈 글씨 데이터에 과적합 되는 것을 방지하고 데이터를 증강하기 위한 목적으로 사용된다(Fig. 12). 실험에선 기본 학습 배치와 참조 데이터가 포함된 배치가 교대로 입력된다. 평가 배치는 평가 데이터와 학습된 모든 클래스에 대한 참조 데이터로 구성된다(Fig. 13). 평가 데이터는 각 참조 데이터와의 유사도가 측정되며, 그 값이 가장 큰 클래스를 평가 데이터의 클래스라 판단한다.

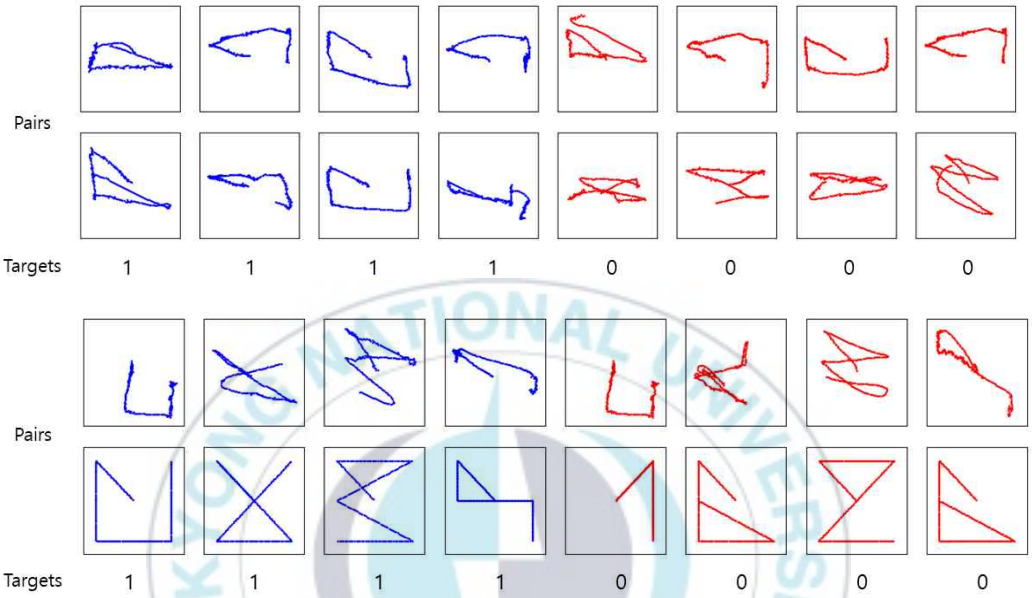

Fig 12. A basic train batch with only eye written characters(top) and a batch with reference data(bottom) when batch size is 8

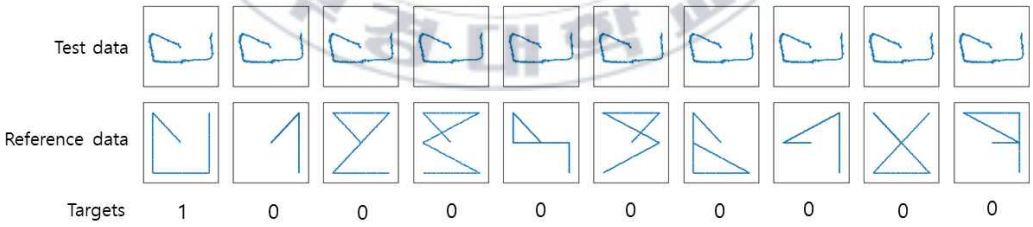

Fig 13. A test batch on eye writing characters of Arabic numbers when the target class is 0

3. 실험 및 결과

제 3장에선 본 연구의 실험과 그 결과에 대해 세부적으로 소개한다. 실험은 총 4가지로 순차적으로 구성된다(Fig. 14). 실험 간 연관성이 존재하며 각 실험 단계는 이전 실험의 주요한 결과를 반영한다. 1) 첫 번째 실험에서는 데이터의 전처리에 주목하여 최적의 전처리 과정을 탐색한다. 아라비아 숫자 데이터를 활용하며, 눈 깜빡임 제거 과정을 수행한 데이터 셋과 수행하지 않은 데이터 셋을 각각 생성한다. 생성된 두 데이터에 따른 샴 네트워크의 성능 차이를 바탕으로 눈 깜빡임 제거의 불필요성을 증명한다. 2) 두 번째 실험은 ViT 특징 추출기와 CNN+LSTM 특징 추출기의 성능을 비교한다. 이때 실험에 사용한 전처리 과정은 이전의 실험 결과를 바탕으로 선정된다. 해당 실험은 ViT의 최적 하이퍼파라미터를 탐색한 후, CNN+LSTM 특징 추출기와 비교하여 우수한 성능의 특징 추출기를 선정한다. 특징 추출기 선정이 완료되면 아라비아 숫자 데이터와 가타카나 문자 데이터의 인식 성능을 확인한다. 3) 세 번째 실험에선 아라비아 숫자 데이터와 가타카나 문자 데이터 모두 학습한 샴 네트워크의 성능을 확인한다. 이때 전처리 과정 선정과 특징 추출기는 이전 실험들을 통해 선정된 우수한 성능의 모델과 최적 파라미터를 사용한다. 4) 마지막 실험은 학습 클래스 증가와 특징 추출기의 개선에 따른 학습되지 않은 클래스 인식(zero-shot learning)에 대한 성능 평가를 진행하고 그 결과를 확인한다.

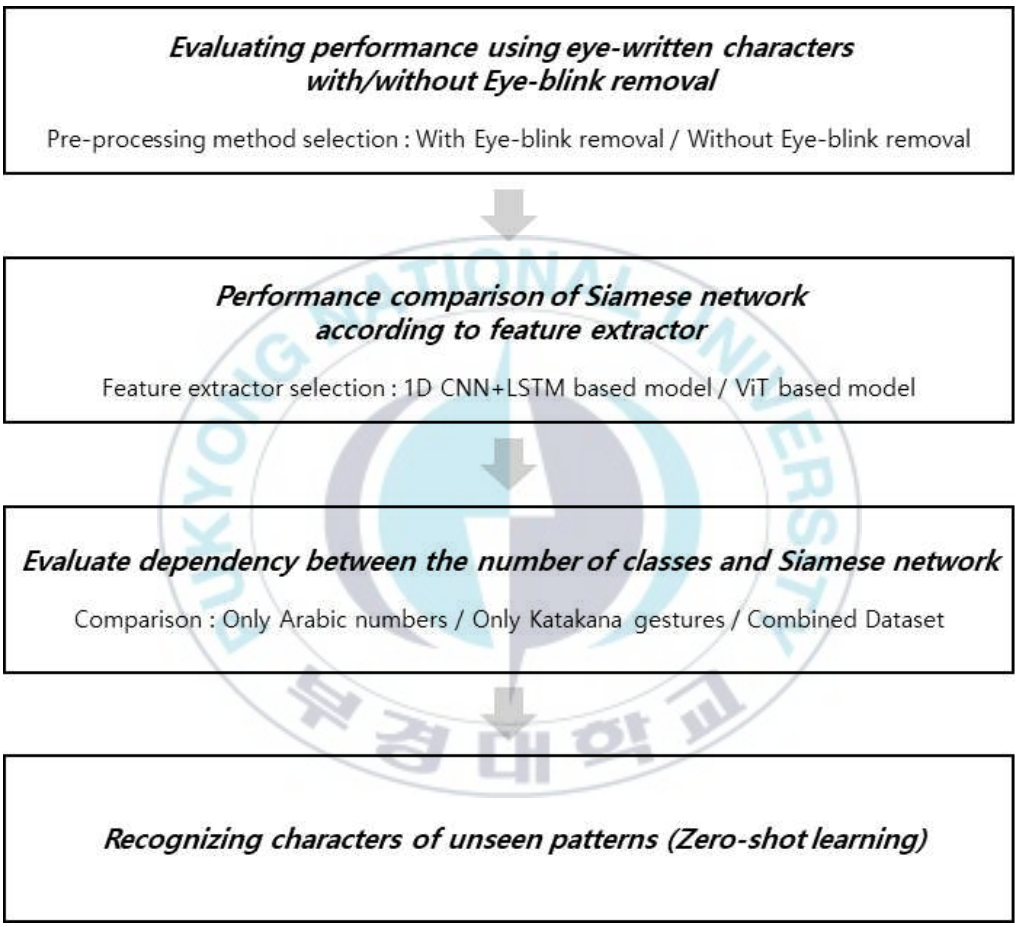

Fig 14. Experiments flow-chart

3.1 실험 절차

본 연구의 모든 실험은 10개의 모델을 생성하여 최적화를 진행한다. 이때 학습 데이터와 평가 데이터는 2:1의 비율로 무작위 선별되어 학습과 평가가 이루어진다. 각 회차의 모델은 학습 시 100 에포크(epoch)가 부여되며 마지막 에포크에서의 분류 정확도를 최적화된 모델의 성능이라 판단한다. 모델의 성능 평가 지표로는 10회 실시된 실험의 평균, 최고, 최저 정확도와 표준편차의 평균으로 표현한다.

실험을 위해 생성된 모델의 최적화 함수는 학습률(learning rate)이 0.001인 아담 함수(Adam optimizer)를 사용하였으며, 손실 함수(loss function)는 이진 크로스 엔트로피(binary cross entropy)가 공통적으로 적용된다.

3.2 눈 깜빡임 제거(eye-blink removal)에 따른 성능 평가

본 실험은 눈 글씨 데이터의 전처리 과정에서 눈 깜빡임 제거에 따른 샴 네트워크의 성능을 비교·분석한다. 아라비아 숫자 데이터를 사용하며 눈 깜빡임 제거 과정을 수행한 데이터와 수행하지 않은 데이터를 개별로 생성한 후 각각의 모델에 입력된다. 이때 샴 네트워크의 특징 추출기는 CNN+LSTM 모델을 사용하며 최적 파라미터 인자들은 저자의 기존 연구를 따른다[15].

Table 1에서 확인할 수 있듯이 눈 깜빡임이 제거된 데이터를 사용한 경우 10회 평균 88.8% 최고 90.6%, 제거되지 않은 데이터의 경우 평균 88.3% 최고 94.3%의 분류 정확도를 달성하였다. 표준편차 또한 눈 깜빡임이 제거되지 않은 경우가 제거된 경우보다 1.205 낮은 우수한 결과가 나타났다.

본 실험의 결과를 미루어보아 샴 네트워크를 활용해 눈 글씨를 인식함에 있어 눈 깜빡임 제거는 약간의 성능 향상을 야기하지만, 공정 시간을 고려하면 불필요한 과정이라고 판단할 수 있다.

Table 1. Comparative results between with and without eye blink removal on Arabic numbers

	Accuracy of 10 times (%)			Std.
	Avg.	Best.	Worst.	
Without eye-blink removal	88.0	90.6	84.3	2.125
With eye-blink removal	88.3	94.3	83.7	3.33

3.3 특징 추출기에 따른 성능 평가

본 실험은 샴 네트워크의 특징 추출기에 따른 분류 정확도를 평가한다. 실험에선 눈 깜빡임이 제거되지 않은 아라비아 숫자 데이터와 가타카나 문자 데이터를 사용한다. 실험은 ViT의 최적 파라미터를 탐색하는 단계와 CNN+LSTM 특징 추출기와의 성능을 비교하는 단계로 구성된다. ViT는 5개의 파라미터를 인자로 입력받으며 Table 2에 해당하는 인자들에 대한 조합으로 격자 탐색(grid search) 후, 최적 파라미터를 선정한다. 다음, 최적 파라미터를 부여받은 ViT와 기존 연구에서 사용한 특징 추출기의 성능을 비교한다.

최적 파라미터가 적용된 ViT 기반 샴 네트워크의 성능은 Table 3에서 확인할 수 있다. 해당 모델은 가타카나 문자 데이터를 학습했을 때 평균 78.0% 최대 84.7%의 분류 정확도를 달성하였다. 아라비아 숫자 데이터를 학습했을 때 평균 88.2% 최대 91.9% 분류 정확도를 달성하였으며, Table 2와 비교하였을 때, CNN+LSTM 기반의 샴 네트워크의 성능과 큰 차이를 보이지 않았다. 하지만, 아라비아 숫자 데이터에 대한 학습 과정에서 ViT 기반 샴 네트워크가 CNN+LSTM 기반 샴 네트워크보다 빠른 최적화를 이룰 수 있다는 것을 Fig. 15을 통해 확인할 수 있다.

구체적으로 Fig. 16에서 클래스별 평가 데이터의 인식 성능을 확인할 수 있다. Fig 16에 따르면 가타카나 문자 12번 클래스는 2,4,9번 클래스로 오인식하는 경향이 나타났으며, 특히 2번 클래스로 치우쳐진 결과를 보인다. 이는 제 2장의 참조 데이터에서 확인할 수 있듯이 두 클래스의 형태가 유사한 것이 원인인 것으로 추측한다.

Table 2. Parameters for ViT feature extractor and Hyperparameters

Parameters	Values		
Hidden size of transformer output	**128**	256	512
Patch size of input data	5		**10**
Number of heads in encoder	4		**8**
Number of encoder blocks	**8**		12
Number of MLP layers and units size	128, 64		**64, 32**

Table 3. Recognition accuracy of ViT based Siamese network using Katakana gestures and Arabic numbers

	Accuracy of 10 times (%)			Std.
	Avg.	Best.	Worst.	
katakana gestures	78.0	84.7	73.1	3.96
arabic numbers	88.2	91.9	85.0	2.62

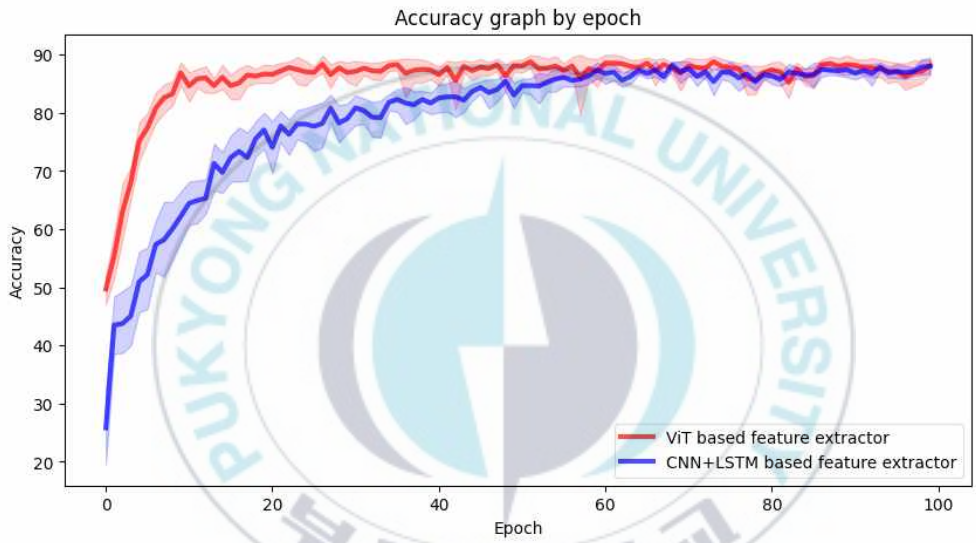

Fig 15. Performance of ViT and CNN+LSTM based Siamese network using Arabic numbers

3.4 클래스 수에 따른 성능 평가

본 실험은 클래스 수의 증가가 샴 네트워크에 미치는 영향을 알아보고자 한다. 데이터 셋은 눈 깜빡임이 제거되지 않은 아라비아 숫자 데이터와 가타카나 문자 데이터를 함께 사용한다. 특징 추출기는 3.2의 실험에서 우수한 성능을 보인 ViT를 사용한다.

실험 결과 평균 79.2% 최대 81% 인식 정확도를 달성한 것을 확인할 수 있다 (Table 4). 구체적으로 Fig. 17에 따르면 클래스와 표본의 증가가 이루어졌음에도 아라비아 숫자와 가타카나 문자의 인식 정확도는 개별적으로 학습한 모델과 유사하다는 것을 확인할 수 있다.

Table 4. Recognition accuracy of ViT based Siamese network using a dataset with Arabic numbers and Katakana gestures combined

	Accuracy of 10 times (%)			Std.
	Avg.	Best.	Worst.	
katakana gestures + arabic numbers	79.2	81.0	78.0	1.07

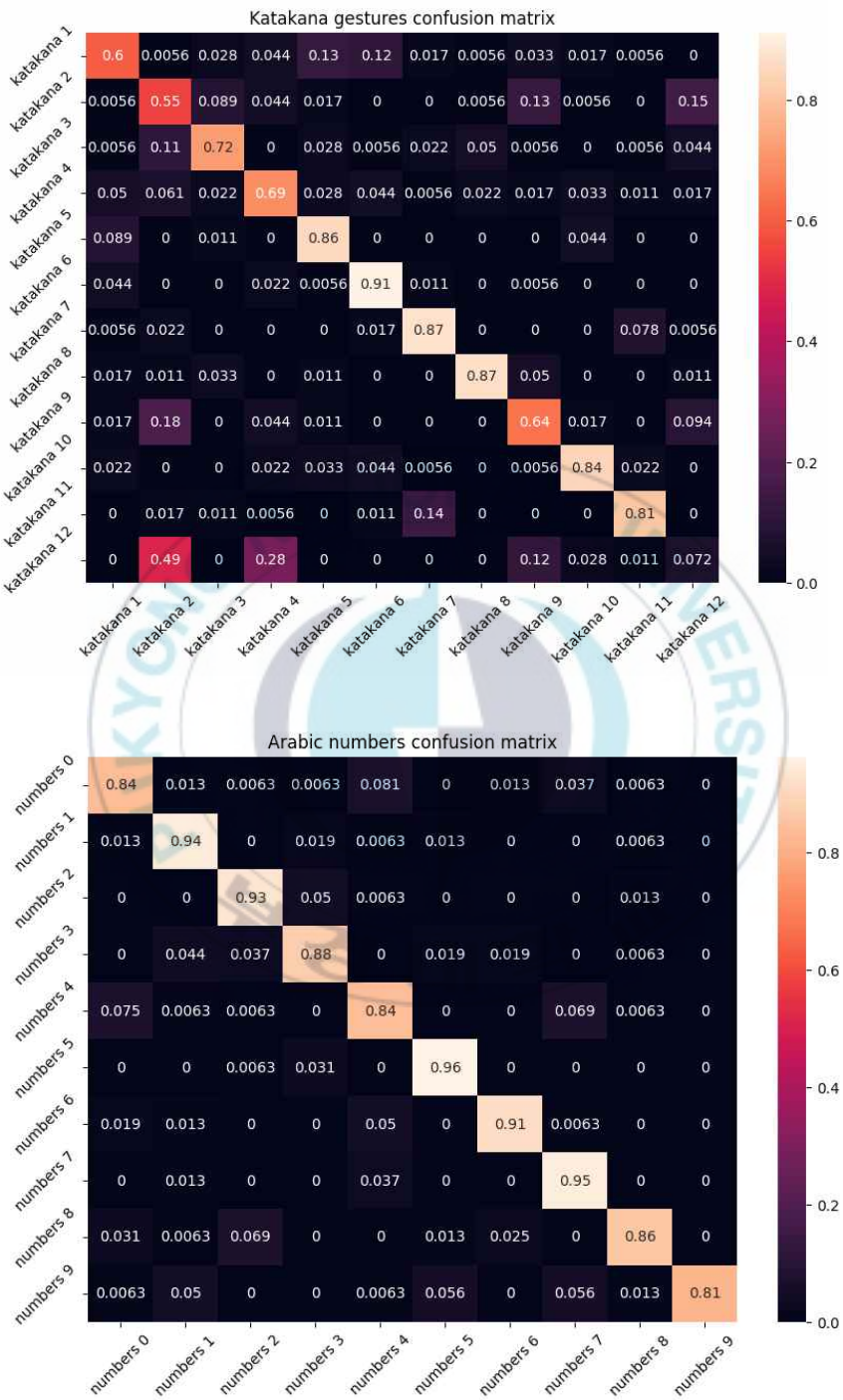

Fig 16. Confusion matrix of individual dataset

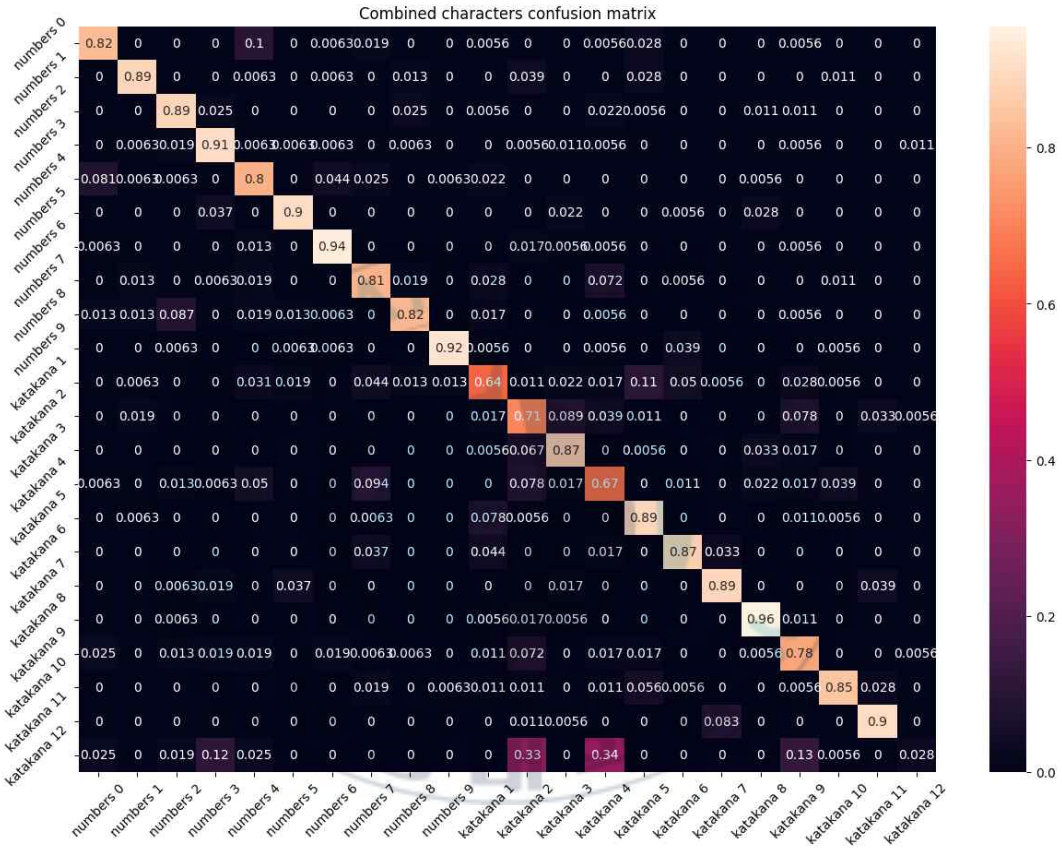

Fig 17. Confusion matrix of dataset with Arabic numbers and katakana gestures combined

3.5 학습되지 않은 클래스 인식 (Zero-shot learning)

본 실험은 샴 네트워크에 학습되지 않은 눈 글씨 형태를 인식하는 실험으로서 Zero-shot learning이라 불린다. 예를 들어, 개와 고양이 이미지를 학습한 샴 네트워크 모델로 얼룩말과 소 이미지를 인식하는 과정이다. Koch 등의 연구[16]에서 샴 네트워크는 특정 이미지들의 일치 여부를 충분히 구분함과 동시에 학습되지 않은 이미지 또한 인식하는 것을 증명하였다. 하지만, 눈 글씨 인식에 샴 네트워크를 적용한 기존의 연구[15]에선 유의미한 결과가 나타나지 않았다. 이는 적은 수의 클래스와 데이터 표본을 사용해 실험을 한 것을 원인으로 판단하였다. 따라서 본 실험에선 기존의 연구와 달리 아라비아 숫자 데이터와 가타카나 문자 데이터를 사용해 표본과 클래스 수를 증가시킴으로써 성능을 개선하고자 한다.

구체적으로 샴 네트워크는 특정 클래스를 제외한 나머지 클래스를 학습한 후, 제외된 특정 클래스에 해당하는 모든 데이터의 클래스를 예측한다. 예를 들어, 아라비아 숫자 0번 클래스를 Zero-shot learning하기 위해 샴 네트워크는 아라비아 숫자 0번을 제외한 아라비아 숫자 1-9 클래스와 가타카나 획 데이터의 모든 클래스를 학습한 후, 아라비아 숫자 0번 데이터에 대한 인식 정확도를 평가한다 (Fig. 18). 본 실험에선 22종의 클래스(아라비아 숫자 10종 + 가타카나 획 12종)에 대해 개별적으로 샴 네트워크 모델을 생성하고 Zero-shot learning를 순차적으로 수행한다. 이때 모델에 사용된 특징 추출기를 최적 파라미터가 적용된 ViT를 사용한다.

실험의 결과는 Table 5에서 확인할 수 있다. Table 5에 따르면 아라비아 숫자 클래스 0과 가타카나 문자 클래스 12를 제외하곤 대부분 평균 정확도 90% 달성하였다. Fig. 19에 따르면 아라비아 숫자 데이터만을 학습하였을 때 숫자 0번 클래스는 높은 인식률을 달성하였지만 Zero-shot learning에서는 4번 클래스로 오인식하는 경향이 나타났다. 또한, 가타카나 문자 12번 클래스는 가타카나 문자 2,4,9번 클래스로 인식하며 실험 3.2에서의 결과와 유사하다.

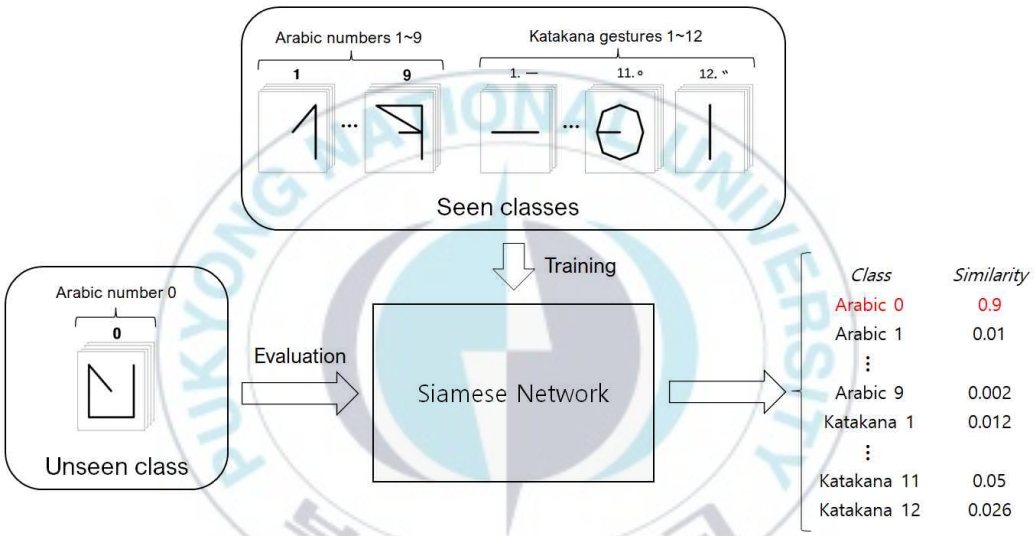

Fig 18. Illustration of Zero-shot learning when the unseen class is Arabic number 0

Table 5. Performances of Zero-shot learning inference with unseen classes

	Accuracy of 10 times (%)			Std.
	Avg.	Best.	Worst.	
numbers 0	21.4	35.2	5.6	11.186
numbers 1	97.4	100.0	94.4	2.512
numbers 2	97.0	100.0	94.4	1.889
numbers 3	97.4	100.0	96.3	1.481
numbers 4	88.9	94.4	83.3	3.704
numbers 5	97.4	98.1	96.3	0.907
numbers 6	97.8	100.0	94.4	2.16
numbers 7	88.9	100.0	68.5	10.798
numbers 8	96.3	100.0	92.6	2.619
numbers 9	95.9	98.1	92.6	2.16
katakana 1	89.3	93.3	86.7	2.261
katakana 2	90.7	96.7	85.0	4.295
katakana 3	97.7	100.0	96.7	1.333
katakana 4	92.3	95.0	90.0	1.7
katakana 5	92.0	96.7	85.0	4.137
katakana 6	93.0	98.3	73.3	9.855
katakana 7	98.0	98.3	96.7	0.667
katakana 8	98.3	100.0	95.0	1.826
katakana 9	89.0	98.3	78.3	7.118
katakana 10	96.3	100.0	93.3	2.211
katakana 11	99.0	100.0	96.7	1.333
katakana 12	28.0	50.0	1.6	18.361

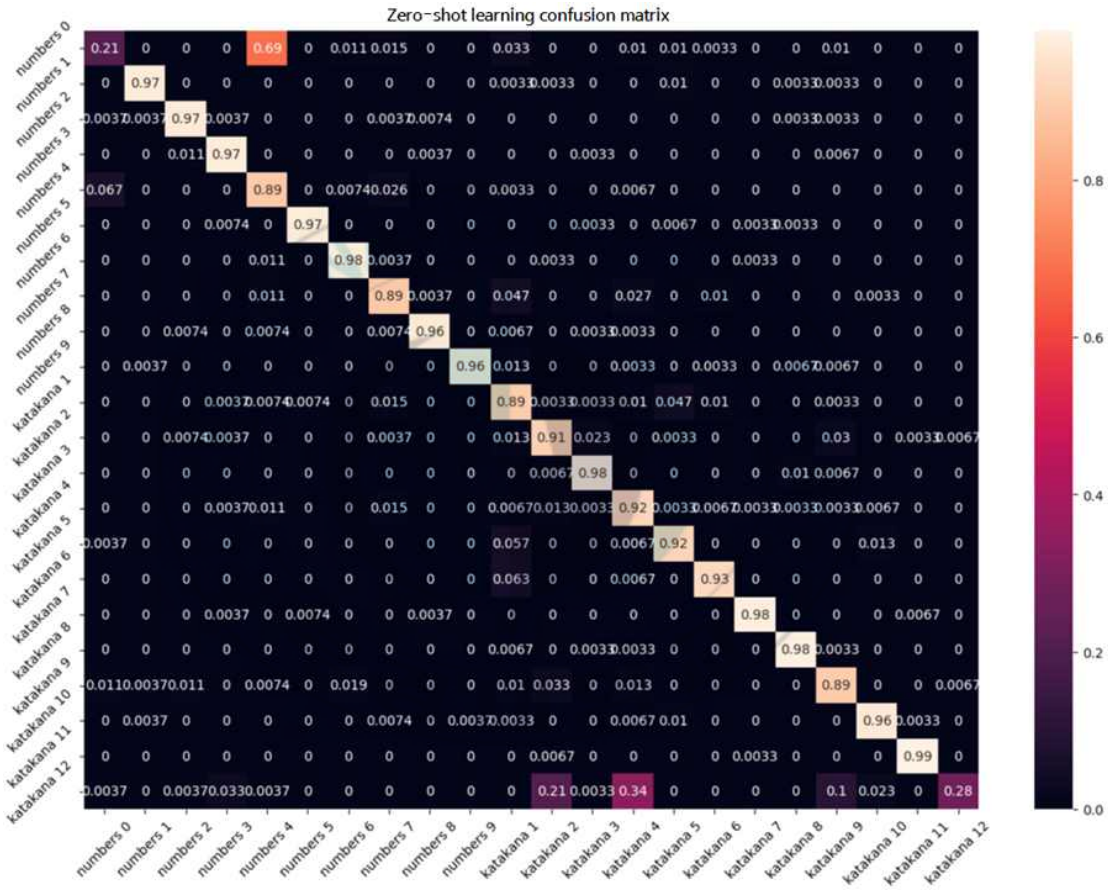

Fig 19. Confusion matrix of Zero-shot learning with unseen classes

4. 결론 및 토의

4.1 결론

본 연구는 샴 네트워크를 활용한 눈 글씨 인식에 대한 연구[15]를 네 가지 실험을 순차적으로 수행해 보완하였다.

첫 번째 실험은 데이터 전처리 관점에서 눈 깜빡임 제거와 같은 불필요한 과정을 최소화하는 것에 초점을 두었다. 실험 3.2의 결과에 따르면 눈 깜빡임이 제거된 데이터는 제거되지 않은 데이터에 비해서 평균 +0.3%p의 인식률 차이를 보였다. 이는 눈 글씨를 인식함에 있어 눈 깜빡임 제거는 약간의 성능 향상을 야기하지만, 전·후처리 시간을 고려하면 그 효과는 미미한 것을 증명하였다.

두 번째 실험은 샴 네트워크의 특징 추출기에 따른 인식률을 평가하였다. 본 연구에서 사용된 특징 추출기는 시계열 데이터에 적합하게 계량된 ViT와 CNN+LSTM 모델을 사용하여 비교·분석하였다. 실험 3.3에 따르면 최적화된 ViT는 CNN+LSTM 모델과 유사한 성능을 가지지만, 빠른 최적화를 이룰 수 있다는 것을 확인할 수 있었다.

세 번째 실험에서는 눈 글씨 클래스의 증가와 샴 네트워크 성능의 상관관계를 분석해보았다. 해당 실험에선 계량된 ViT 기반 샴 네트워크를 사용하였으며, 아라비아 숫자와 가타카나 문자 데이터를 함께 학습하였다. 결과적으로 평균 79.2% 최대 81% 인식 정확도를 달성하였다. Fig.16의 아라비아 숫자만을 학습한 모델의 성능을 표현한 혼동 행렬과 Fig.17에서 가타카나 획 데이터와 함께 학습된 모델의 성능을 표현한 혼동 행렬 중 아라비아 숫자 데이터에 대한 수치를 비교하면, 샴 네트워크는 클래스의 수에 관계없이 일정한 성능을 유지하는 것을 알 수 있었다. 카타카나 문자 데이터에 대한 수치 또한 동일하였다. 이러한 결과는 학습 클래스 수는 샴 네트워크의 성능에 주는 영향이 적다는 것을 증명한다. 하지만, 형태가 유사한

클래스가 학습될 경우 오인식률이 증가할 것으로 예상한다.

본 연구의 마지막 실험에서는 학습되지 않은 클래스에 대한 인식 정확도 평가가 진행되었다. 해당 실험은 특정 클래스를 제외한 나머지 클래스를 학습한 샴 네트워크에 제외된 특정 클래스의 올바르게 인식하는지에 대한 평가를 실시하였다. Table 4에 따르면 샴 네트워크는 아라비아 숫자 클래스 0(21%)과 가타카나 문자 클래스 12(28%)를 제외한 대부분의 클래스는 90% 이상의 정확도를 달성하는 것을 확인할 수 있다. 아라비아 숫자 클래스 0과 가타카나 문자 클래스 12의 경우 Fig. 19에서 확인할 수 있듯이 샴 네트워크에 사전 학습된 유사한 형태를 가진 아라비아 숫자 4번 클래스, 카타카나 2, 4, 9번 클래스로 인식한다.

이와 같은 결과는 사전학습 된 샴 네트워크에 새로운 클래스의 추가가 필요하다면, 참조 데이터만을 추가함으로써 높은 성능을 달성할 수 있다. 이는 모델의 재학습과 데이터 수집에 대한 비용을 크게 절감하며 생산성 향상에 크게 기여한다.

4.2 연구의 한계 및 향후 연구 방향

본 연구의 한계와 이를 보완할 향후 연구 방향은 크게 세 가지가 있다. 우선, 샴 네트워크는 안구전도로 수집된 눈 글씨 인식에 실용적인 방법이라는 것을 증명하기 위한 실험을 진행하였다. 실험을 통해 제안된 최적 모델은 아라비아 숫자 데이터를 최대 91.9% 평균 88.2%, 가타카나 획 데이터를 최대 84.7% 평균 78.0%의 인식 정확도를 달성하였다. 하지만, 이러한 결과가 동일한 조건에서의 기존의 방법들과 비교하여 우수한지에 대한 여부를 증명하기엔 본 연구에선 부족함이 있다. 따라서 향후의 연구에선 기존의 방법론들과의 비교·분석을 통해 모델의 성능에 대한 검증과 개선이 필요하다.

본 연구는 클래스 수에 따른 샴 네트워크의 성능 차이를 평가하였다. 실험 결과에 따르면 클래스 수가 증가함에도 개별 클래스에 대한 인식률의 변화는 크지 않았다. 이는 클래스의 수와 표본이 증가함에도 안정적인 성능을 달성할 수 있다는 점에서 의미가 있다. 다만, 유사한 형태의 클래스가 학습될 경우 모델이 성능이 감소 될 것으로 예상한다. 향후 연구에서는 이러한 문제에 초점을 두어 성능 변화가 발생하는 경우와 해결 방안을 탐구하는 다양한 실험이 필요할 것으로 보인다.

마지막으로, 본 연구는 오프라인으로 수집된 데이터를 활용해 모델을 학습하였으며, 모델 평가 또한 오프라인 데이터를 입력하였다. 실험의 결과 준수한 인식 정확도를 달성한 것을 확인하였지만, 실제 트래픽이 발생하는 실시간 환경에서의 실용적인 측면에서의 평가는 이루어지지 않았다. 이러한 결과를 바탕으로 상용화 단계에 이르기 위해선 제안된 모델을 활용한 시스템에 대한 탐구가 필수적이다. 이에 향후의 연구에선 어플리케이션의 확장을 통해 시스템 차원에서의 성능 평가가 이루어질 것을 기대한다.

4.3 연구의 시사점

본 연구는 이미지와 자연어처리 분야에서 사용하던 트랜스포머(transformer) 모델을 시계열 데이터 분석에 적용했다는 점에서 학술적 의의가 있다. 구체적으로 샴 네트워크의 특징 추출기를 ViT를 적용하였으며, 저자의 이전 연구에서 사용된 CNN+LSTM 기반 특징 추출기에 비해 빠른 최적화를 이루어내는 효과를 불러왔다. 이는 생체 신호와 같은 시계열 데이터 분석에 있어 트랜스포머 모델은 강력한 대안이 될 수 있다는 것을 증명하였다.

또한, 제한된 데이터 표본으로 의미있는 성능을 달성할 수 있다. 안구전도 기반 눈 글씨 인식에 가장 큰 한계는 활용 가능한 데이터의 표본이 부족하다는 것이다. 본 연구를 통해 적은 데이터만으로도 높은 인식률을 달성할 수 있다는 것을 증명하였으며 대안점을 제시하였다. 이는 데이터 수집에 대한 비용 절감에 큰 기여를 할 것이라 기대한다.

눈 글씨를 인식하는 다양한 연구에선 학습되지 않은 눈 글씨 데이터를 인식할 수 없다는 한계가 명확하였다. 한편, 본 연구에선 샴 네트워크를 통해 학습되지 않은 클래스 또한 높은 인식률을 달성할 수 있다는 것을 증명하였다. 이는 학습 클래스가 추가될 때마다 주기적인 재학습이 필요하다는 기존의 한계를 해결하였으며, 결론적으로 새로운 클래스를 추가할 때 데이터 수집에 의존하지 않아도 된다는 점에서 의미를 가진다.

References

[1] Esposito, Daniele, et al. "Biosignal-based human–machine interfaces for assistance and rehabilitation: A survey." Sensors 21.20 (2021): 6863.

[2] Belkhiria, Chama, et al. "EOG-Based Human–Computer Interface: 2000–2020 Review." Sensors 22.13 (2022): 4914.

[3] Malmivuo, Jaakko, and Robert Plonsey. Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields. Oxford University Press, USA, 1995.

[4] Chang WD. Electrooculograms for Human-Computer Interaction: A Review. Sensors (Basel). 2019 Jun 14;19(12):2690. doi: 10.3390/s19122690. PMID: 31207949; PMCID: PMC6630230.

[5] Yamagishi K, Hori J, Miyakawa M. Development of EOG-based communication system controlled by eight-directional eye movements. Conf Proc IEEE Eng Med Biol Soc. 2006;2006:2574-7. doi: 10.1109/IEMBS.2006.259914. PMID: 17945724.

[6] Kwang-Ryeol Lee, Won-Du Chang, Sungkean Kim, Chang-Hwan Im. Real-Time "Eye-Writing" Recognition Using Electrooculogram. IEEE Trans Neural Syst Rehabil Eng. 2017 Jan;25(1):37-48. doi: 10.1109/TNSRE.2016.2542524. Epub 2016 Mar 15. PMID: 28113859.

[7] Hosni SM, Shedeed HA, Mabrouk MS, Tolba MF. EEG-EOG based Virtual Keyboard: Toward Hybrid Brain Computer Interface. Neuroinformatics. 2019 Jul;17(3):323-341. doi: 10.1007/s12021-018-9402-0. PMID: 30368637.

[8] Choi, Jin Woo, and Sungho Jo. “Implementation of EOG Signal to Chess Game Control.” KIISE Transactions on Computing Practices, vol. 24, no. 8, Korean Institute of Information Scientists and Engineers, 31 Aug. 2018, pp. 416 - 421. Crossref, doi:10.5626/ktcp.2018.24.8.416.

[9] Tsai, Jang-Zern et al. “A Feasibility Study of an Eye-writing System Based on Electro-oculography.” Journal of Medical and Biological Engineering 28 (2008): 39-46.

[10] K. -R. Lee, W. -D. Chang, S. Kim and C. -H. Im, "Real-Time “Eye-Writing” Recognition Using Electrooculogram," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 1, pp. 37-48, Jan. 2017, doi: 10.1109/TNSRE.2016.2542524.

[11] Chang, W. D., Cha, H. S., Kim, D. Y., Kim, S. H., & Im, C. H. (2017). Development of an electrooculogram-based eye-computer interface for communication of individuals with amyotrophic lateral sclerosis. Journal of neuroengineering and rehabilitation, 14(1), 89. https://doi.org/10.1186/s12984-017-0303-5

[12] Fang, Fuming, and Takahiro Shinozaki. "Electrooculography-based continuous eye-writing recognition system for efficient assistive communication systems." PloS one 13.2 (2018): e0192684.

[13] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[14] Chang, W.-D.; Choi, J.-H.; Shin, J. Recognition of Eye-Written Characters Using Deep Neural Network. Appl. Sci. 2021, 11, 11036. https://doi.org/10.3390/app112211036

[15] Kang, D.-H.; Chang, W.-D. Recognition of Eye-Written Characters with Limited Number of Training Data Based on a Siamese Network. Electronics 2021, 10, 3009. https://doi.org/10.3390/electronics10233009

[16] Koch, Gregory R.. “Siamese Neural Networks for One-Shot Image Recognition.” (2015).

[17] Chang, W.-D.; Cha, H.-S.; Kim, K.; Im, C.-H. Detection of eye blink artifacts from single prefrontal channel electroencephalogram. Comput. Methods Programs Biomed. 2016, 124, 19 - 30.

[18] Bromley, Jane, et al. "Signature verification using a" siamese" time delay neural network." Advances in neural information processing

systems 6 (1993).

[19] Chopra, Sumit, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.

[20] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).