The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 23, No. 3, pp.145-151, Jun. 30, 2023. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2023.23.3.145
JIIBC 2023-3-20

3D-CNN에서 동적 손 제스처의 시공간적 특징이 학습 정확성에 미치는 영향

Effects of Spatio-temporal Features of Dynamic Hand Gestures on Learning Accuracy in 3D-CNN

정영지*

Yeongjee Chung*

요 약 3D-CNN은 시계열 데이터 학습을 위한 딥 러닝 기법 중 하나이다. 이러한 3차원 학습은 많은 매개변수를 생성할 수 있으므로 고성능 기계학습이 필요하거나 학습 속도에 커다란 영향을 미칠 수 있다. 본 연구에서는 손의 동적인 제스처 동작을 시공간적으로 학습할 때, 3D-CNN 모델의 구조적 변화 없이 입력 영상 데이터의 시공간적 변화에 따른 학습 정확성을 분석함으로써, 3D-CNN을 이용한 동적 제스처 학습의 효율성을 높이기 위한 입력 영상 데이터의 최적 조건을 찾고자 한다. 첫 번째로 동적 손 제스처 영상 데이터에서 동적 이미지 프레임의 학습구간을 설정함으로써 제스처 동작간 시간 비율을 조정한다. 둘째로는 클래스간 2차원 교차 상관 분석을 통해 영상 데이터의 이미지 프레임간 유사도를 측정하여 정규화 함으로써 프레임간 평균값을 얻고 학습 정확성을 분석한다. 이러한 분석을 통하여, 동적 손 제스처의 3D-CNN 딥 러닝을 위한 입력 영상 데이터를 효과적으로 선택하는 두 가지 방법을 제안한다. 실험 결과는 영상 데이터 프레임의 학습구간과 클래스간 이미지 프레임간 유사도가 학습 모델의 정확성에 영향을 미칠 수 있음을 보여준다.

Abstract 3D-CNN is one of the deep learning techniques for learning time series data. Such three-dimensional learning can generate many parameters, so that high-performance machine learning is required or can have a large impact on the learning rate. When learning dynamic hand-gestures in spatiotemporal domain, it is necessary for the improvement of the efficiency of dynamic hand-gesture learning with 3D-CNN to find the optimal conditions of input video data by analyzing the learning accuracy according to the spatiotemporal change of input video data without structural change of the 3D-CNN model. First, the time ratio between dynamic hand-gesture actions is adjusted by setting the learning interval of image frames in the dynamic hand-gesture video data. Second, through 2D cross-correlation analysis between classes, similarity between image frames of input video data is measured and normalized to obtain an average value between frames and analyze learning accuracy. Based on this analysis, this work proposed two methods to effectively select input video data for 3D-CNN deep learning of dynamic hand-gestures. Experimental results showed that the learning interval of image data frames and the similarity of image frames between classes can affect the accuracy of the learning model.

Key Words : 2D-Cross correlation, 3D-CNN, Dynamic Hand-Gesture Recognition, Human-Computer Interface, patiotemporal Features

*정회원, 원광대학교 컴퓨터·소프트웨어공학과
접수일자 2023년 4월 21일, 수정완료 2023년 5월 21일
게재확정일자 2023년 6월 9일

Received: 21 April, 2023 / Revised: 21 May, 2023 / Accepted: 9 June, 2023
*Corresponding Author: yjchung@wku.ac.kr
Dept. of Computer·Software Engineering, Wonkwang University, Korea

Ⅰ. 서 론

Imaging Net Challenge[1] 에서의 Alex Net의 기록적인 성능 향상 이후 많은 연구자들이 2D-CNN(2D Convolutional Neural Network)구조를 여러 애플리케이션과 인간·컴퓨터 상호작용 분야에 적용하기 시작했다. 대표적으로 2D-CNN 구조를 그대로 가져와 각 이미지 프레임에 적용하려는 시도가 있었다. 이러한 시도는 신경망에서 구조적으로 시공간적인 정보를 활용하지 못하기 때문에 제한적이었다. 따라서 이러한 구조적 문제를 해결하기 위해서는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Networks)의 조합으로 특수 기능을 CNN으로 학습하고 이를 LSTM(Long Short-Term Memory)으로 학습한다. 그러나 이러한 조합 역시 기존 연구에 비해 성능 면에서 제한적이었고, 컨볼루션 필터가 특수한 특징만 학습하는 구조적 문제는 남아있었다. 이러한 2D-CNN 단일 구조의 한계를 극복하고 더 많은 시공간적 특징을 학습하기 위해 여러 가지 방법이 고안되었으며 그 중에서 시공간적인 특징을 적용할 수 있는 3D 컨볼루션 신경망 모델에서 상당한 성과를 얻을 수 있었다.[13]

3D 컨볼루션 신경망(3D-CNN)은 컨볼루션 필터를 모두 3차원적으로 사용하여 3D 컨볼루션을 얻는 접근 방식이다. 따라서 하나의 필터에 의해 생성된 특징 맵도 3차원이다. 이러한 구조 덕분에 3D-CNN은 컨볼루션 필터 자체에서 연속 프레임의 시공간적 특징을 학습할 수 있다. 이 구조는 단기적인 시공간적 특징 학습을 허용한다.[2] 그러나 3D-CNN 방식은 필터가 3D이고 훨씬 더 많은 파라미터가 있으며, 2D-CNN과 같은 사전 훈련된 모델이 없기 때문에 학습을 진행하기 어려운 단점이 있다.

이 때문에 충분히 깊은 구조를 구축하는 것이 곤란하고 모델의 구조를 변형하여 높은 정확도를 얻는 데 한계가 있다. 따라서 본 연구에서는 3D-CNN의 구조적 변형 없이 동적 제스처의 입력 데이터만 변경하였을 때 모델의 정확도 변화를 분석하고자 한다. 방법으로는 동적 제스처의 입력 데이터의 시공간적 프레임 구간과 클래스 간 2차원 상호 상관에 따른 프레임간 중복성의 두 가지 파라미터에 따라 입력 데이터를 변환하고 정확도의 변화를 분석한다.

Ⅱ. 관련 연구

RNN 알고리즘은 반복적이고 순차적인 데이터 학습에 특화된 인공신경망의 일종으로 내부 순환 구조를 특징으로 한다. 순환 구조를 이용하여 가중치를 통해 과거 학습이 현재 학습에 반영된다. RNN은 기존의 연속적이고 반복적이며 순차적인 데이터 학습의 한계를 해결하였다. 또한 현재 학습과 과거 학습 사이의 연결을 가능하게 하고 시간 의존적이라는 특징이 있다. 주로 음성 파형이나 텍스트의 앞뒤 구성 요소를 식별하는 데 사용되기도 한다.[2-3]

RNN은 해당 정보와 해당 정보가 사용되는 지점 사이의 거리가 크게 줄어들면 역전파(back-propagation) 과정에서 점진적으로 그래디언트를 줄이는 것으로 알려져 있다. 이를 Vanishing Gradient 문제라고 하는데 이를 극복하기 위해 고안된 것이 LSTM이다. LSTM은 RNN의 숨겨진 상태에 셀 상태를 추가한 구조이다. 시간이 지남에 따라 계산되는 RNN의 방법과 유사하지만 숨겨진 계층의 계산 방법에 차이가 있다. LSTM은 네 단계로 구성된다. 1단계는 시그모이드 함수를 사용하여 삭제할 데이터를 선택한다. 2단계는 시그모이드 및 하이퍼볼릭 접선 함수(Tanh)를 사용하여 새 데이터가 LSTM 셀 단계에 저장되는지 여부를 결정한다. 3단계는 셀 상태를 업데이트하고 마지막 단계에서는 Tanh 함수의 출력 값과 시그모이드 함수를 통과한 셀 상태를 결정한다.[4] 이 외에도 2개의 스트림을 사용하는 Two-Stream Net, 2D ConvNet을 3D ConvNet으로 변환하는 Inflated 3D ConvNet(I3D) 등 영상 인식에 대한 여러 연구들이 있다.[5-7,15]

Ⅲ. 3D-CNN에 의한 동적 손 제스처의 학습

1. 3D-CNN 구조

3D-CNN은 비디오 모델링에서 가장 많이 사용되는 모델 중 하나이다. 3D CNN을 사용하면 시공간 정보의 손실 없이 비디오 데이터를 학습할 수 있다. 2D-CNN에 비해 3D-CNN은 3D 컨볼루션 및 3D 풀링 작업으로 시간 정보를 생성한다. 즉, 2D-CNN에서 공간적으로만 수행되는 작업을 3D-CNN에서는 시공간에 걸쳐 수행할 수 있다. 그림 1은 단일 컨볼루션의 개념을 상대적으로 비교한 것이다. 2D 컨볼루션은 여러 영상에 적용해도 하나의 채널로 인식되기 때문에 하나의 영상만 생성되지만 3D 컨볼루션은 입력 신호의 시간 정보를 그대로 유지하고 계산 결과가 시공간적이다.

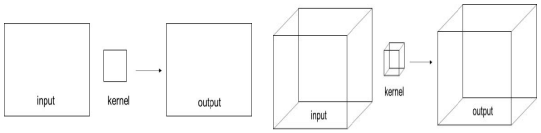


그림 1. 2차원과 3차원 컨볼루션
Fig. 1. 2D and 3D convolution operations.

그림 1(a)는 2차원 컨볼루션의 입력 이미지, 커널 및 출력 형태를 보여주고, 그림 1(b)는 3차원 컨볼루션의 시공간적 입력 이미지, 커널 및 출력 형태를 보여준다.[8-10,14]

2. 데이터 세트

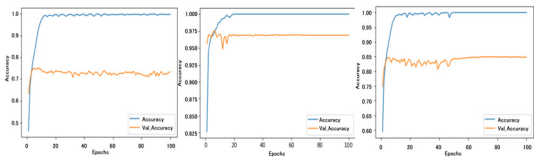
본 연구에서는 200B-Jester-v1 Dataset[11]라는 공개 비디오 클립 데이터 세트를 사용한다. 20BN-Jester -v1 Dataset는 148,092개의 비디오 클립으로 구성되어 있으며 27개 레이블 중 8개만 추출하여 사용한다. 각 클래스에 대해 2000개의 클립이 사용되었으며 각 클립에 대해 30개의 프레임이 추출되었다. 표 1은 우리가 사용한 제스처 목록을 보여준다.

표 1. 제스처 목록
Table 1. Dynamic Hand-Gesture list

클래스 레이블
제스처 없음
두 손가락을 왼쪽으로 밀기
두 손가락을 오른쪽으로 밀기
손바닥 들기
왼쪽으로 밀기
오른쪽으로 밀기
엄지 손가락 들기
엄지 손가락 내리기

3. 동적 손 제스처의 학습

1개 클래스부터 10개 클래스까지 진행한 결과, 2개 클래스의 경우 99%에 가까운 학습 정확도의 증가율을 보였지만, 클래스 수가 늘어날수록 정확도가 떨어진다. 높은 정확도를 보이는 매개 변수를 찾기 위해 클래스 수, 실행 속도 및 배치 크기를 변경하여 제스처를 훈련한다. 먼저, 학습에 사용되는 클래스의 수가 정확도에 얼마나 영향을 미치는지 분석한다. 학습은 2개 클래스, 4개 클래스, 8개 클래스로 나누어 진행한다. 표 2와 그림 2는 모델의 학습 정확도를 보여주고 있다. 클래스 수가 증가함에 따라 학습 정확도가 크게 떨어지는 것을 알 수 있다.



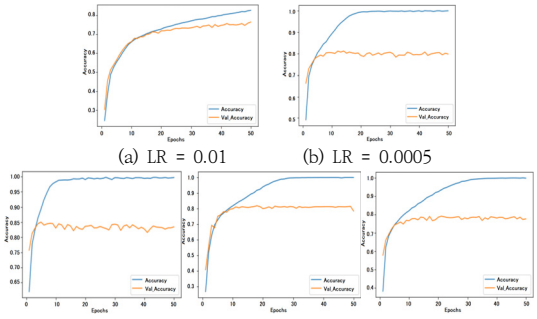
(a) 2 Classes (b) 6 Classes (c) 8 Classes

그림 2. 클래스 수에 따른 학습 정확도
Fig. 2. Learning accuracy by number of Classes

표 2. 클래스 수에 따른 학습 정확성
Table 2. Learning Accuracy by number of Classes

Class 개수	정확성(val_accuracy)
2	0.9688
6	0.8470
8	0.7318

다음으로 학습동안 정확도에 미치는 학습률(Learning Rate: LR)의 영향을 분석한다. LR은 0.01(1%)에서 0.00001(0.001%)까지 5단계로 실험하였다. 그림 3 과 표 3 은 LR에 의한 학습 결과에 따른 학습 정확도를 보여준다. 그림 3(b)는 가장 높은 정확도를 보이는 학습 결과를 보여주고, 그림 3의 (e)는 가장 낮은 정확도를 보이는 학습 결과를 보여주고 있다. 그림 3에서 LR이 증가함에 따라 정확도가 증가하는 것처럼 보이지만, LR = 0.0005에서 LR = 0.01로 증가할 때 역으로 정확도가 감소하는 것을 보여준다.



(a) LR = 0.01 (b) LR = 0.0005
(c)LR=0.0001 (d)LR=0.00005 (e)LR=0.00001

그림 3. 학습률에 따른 학습 정확도
Fig. 3. Learning accuracy by Learning Rate

표 3. 학습률에 따른 정확성
Table 3. Learning accuracy by learning rate

학습률	정확성(val_accuracy)
0.01(1%)	0.7843
0.0005(0.05%)	0.8333
0.0001(0.01%)	0.7975
0.00005(0.005%)	0.7760
0.00001(0.001%)	0.7632

마지막으로 뱃치 처리하는 크기가 정확도에 미치는 영향을 분석한다. 처리하는 크기가 64일 때 가장 높은 정확도를 보였는데, 메모리 부족(Out of Memory: OOM) 오류를 피하기 위해 처리하는 뱃치 크기를 128까지만 실험하였다. 표 4 및 그림 4는 처리 크기 별 학습 결과를 보여준다.

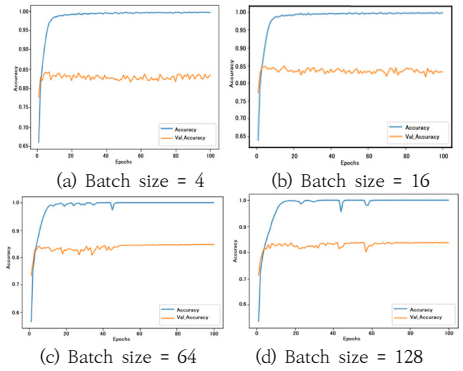


그림 4. 뱃치 처리 크기에 따른 정확도
Fig. 4. Accuracy by Batch size

표 4. 뱃치 처리 크기에 다른 학습 정확도
Table 4. Learning Accuracy by Batch size

처리 크기(batch)	정확성(val_accuracy)
8	0.8345
16	0.8312
64	0.8470
128	0.8363

Ⅳ. 학습을 위한 입력 데이터의 선택

3D-CNN의 구조적 변형 없이 구조의 아무런 변형 없이, 두 가지 측면에서 모델의 정확도에 미치는 영향을 고려한다. 첫 번째는 동적 입력 데이터의 시공간적 프레임의 구간과, 두 번째는 클래스 별로 동적 시공간 데이터의 2차원 교차 상관(2D-Correlation)에 따른 클래스 간 유사성의 두 가지 파라미터에 따라 입력 데이터를 변환하여 학습함으로써 학습 정확성의 변화를 분석한다.

1. 동적 입력 데이터의 시공간적 프레임의 구간

20BN-Jester-v1 Dataset의 경우 각 클래스는 비디오 클립과 프레임으로 구성된다. 프레임은 모든 클립이 그림 5와 같은 구조를 따를 때, [멈춤(Stops), 동작(Actions), 멈춤(Shift-Stops)] 구간으로 나뉜다. 본 연구에서는 그림 5 와 같이 세 가지 방법으로 n개 프레임에서 입력 구간을 선택한다. 입력 프레임은 멈춤 프레임과 동작 프레임이 절반인 중앙 프레임에서 양쪽으로 30프레임과 멈춤 구간이 많은 전방 30프레임과 후방 30프레임으로 입력 구간을 선택한다. 학습 정확도는 영상의 중앙 30프레임 구간에서 가장 높게 나타났으며 영상의 앞, 뒤 구간은 중간 구간보다 정확도가 떨어지는 것으로 나타난다.

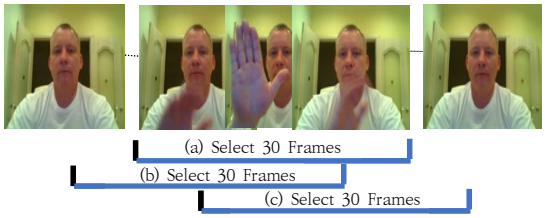


그림 5. 입력 데이터 구간 선택
Fig. 5. Input data interval selection

이는 단일 학습동안 선택된 동적 시공간 입력 데이터에서 어떻게 멈춤 및 동작 구간이 나타나는지에 따라 정확도가 달라질 수 있음을 의미한다. 멈춤 제스처 구간과 동작 제스처 구간이 유사하게 입력 데이터 구간을 선택한 경우에는 정확성이 높고, 하나의 제스처 구간이 다른 제스처 구간보다 크게 입력 데이터 구간을 선택한 경우에는 정확도가 낮다는 것을 확인할 수 있다. 예를 들어 선택한 30프레임 중 멈춤 구간이 15프레임이고 동작 구간이 15프레임이면 멈춤 구간이 1프레인이고 동작 구간이 29프레임인 경우보다 학습 정확성이 높아진다. 이는 3D-CNN모델을 적용하여 동적인 시공간적 데이터를 학습할 때, 입력 데이터의 시작점 선택에 따라 학습모델의 정확성에 영향을 미친다는 것을 의미한다.

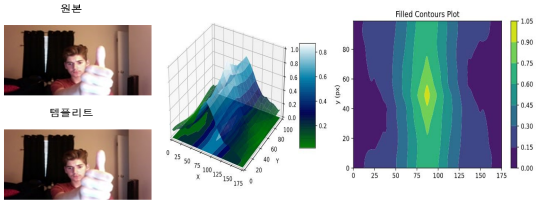
2. 동적 시공간 데이터의 2차원 교차 상관에 따른 클래스 간 유사성

시공간 데이터 클래스 간의 유사도 분석을 위하여, 2차원 교차 상관을 통해 클래스 간의 유사성을 측정하고, 측정된 값을 정규화 한다.[12] 그림 6의 (a), (b), (c)는 각각 동일한 클래스의 입력 데이터 프레임, 유사한 클래스의 입력 데이터 프레임, 상이한 클래스의 입력 데이터 프레임에 대해서 2차원 상호 교차 상관을 측정하여 나타낸 클래스 별 프레임 이미지와 3차원 및 2차원 등고선 매핑 그래프(Contour Mapping Graph)이다. 2D 교차 상관은 식(1)로 얻을 수 있다.

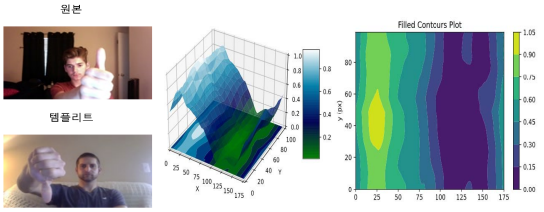
$$G[i,j]=\sum_{u=-k}^{k}\sum_{v=-k}^{k}h[u,v]\,F[i+u,j+v] \quad (1)$$

식(1)에서 i 는 프레임의 폭, j 는 프레임의 높이이다. u와 v는 커널 h의 좌표 값이다. h[u, v]는 선형 조합의 가중치이다. 수식의 기준점이 이미지의 중심이기 때문에 k 값에 음수가 존재한다. 정규화된 2차원 교차 상관은 식(2)로 얻을 수 있다.

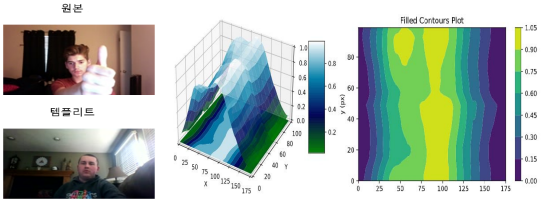
$$g[i,j]_{Norm}=\frac{g[i,j]-g[i,j]_{\min}}{g[i,j]_{\max}-g[i,j]_{\min}} \quad (2)$$



(a) 동일한 클래스의 입력 데이터 프레임



(b) 유사한 클래스의 입력 데이터 프레임



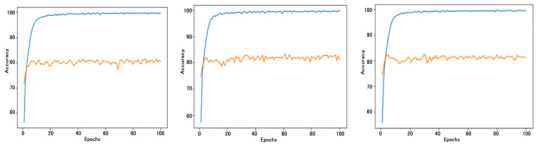
(c) 상이한 클래스의 입력 데이터 프레임

그림 6. 상호 교차 상관을 측정하여 나타낸 3차원 및 2차원 등고선 매핑 그래프
Fig. 6. Contour Mapping Graph for 2D-Cross Correlation of input data frames

입력 데이터에서 클래스 별 해당 인덱스의 프레임 간 2차원 교차 상관을 식(1), 식(2)로 계산하여 입력 데이터의 클래스간 유사도를 측정한다. 또한 다른 매개변수는 고정하고 입력 데이터의 클래스 간 유사도를 측정한 후, 해당 입력 데이터의 프레임으로 학습하였을 때 클래스 간의 유사성이 높을수록 학습 정확성이 낮아지는 것을 확인할 수 있다.

V. 동적 손 제스처 학습 실험 결과

VI.1에서 제시한 방법으로 입력 데이터를 변환한 경우, 동적 시공간 데이터의 중간 30프레임 구간에서 가장 높은 학습 정확성이 측정되었다. 이는 동적 시공간 데이터의 학습에서 동작 구간과 멈춤 구간 비율이 학습 정확도에 영향을 미친다는 것을 의미한다. 이 실험에서 각 구간을 구성하는 프레임 수의 차이가 증가할 때 학습 정확도가 감소하고, 그 정확성의 차이는 약 4%로 측정되었다. 표 5와 그림 6은 시공간 데이터의 입력 선택 구간에 따른 학습 정확도를 보여주고 있다.



(a)Fr.30 Fr. (b)Mid.30 Fr. (c)Rear.30 Fr.

그림 7. 프레임 구간에 따른 학습 정확도
Fig. 7. Accuracy by frame intervals

표 5. 프레임 구간에 따른 정확성
Table 5. Accuracy by frame interval

프레임 선택 구간	정확성(val_accuracy)
전방 30프레임	0.8138
중간 30프레임	0.8470
후방 30 프레임	0.8113

VI.2의 방법으로 세 가지 유형의 동적 손 제스처 세트를 설정하여 학습하고 그 정확도를 측정하였다. 동적 손 제스처 세트는 클래스 간 높은 유사성, 중간 유사성과 낮은 유사성으로 구분하였다. 표 6은 손 제스처 클래스 세트의 이미지 유사도에 따른 학습 정확도를 보여준다. 학습한 결과, 클래스 간 유사도가 높은 클래스는 클래스 간 유사도가 낮은 클래스에 비해 학습 정확도가 미세하게 감소함을 보여주고 있다. 또한, 클래스간 프레임의 유사도가 중간과 높음에서는 정확성에 유의미한 차이가 없었고, 유사도가 낮을수록 학습 정확성이 높다는 것을 알 수 있다.

표 6. 손 제스처 클래스 세트의 유사도에 따른 학습 정확성
Table 6. Learning accuracy by Hand Gesture Class Data set by Similarity

제스처 클래스 세트	유사성	정확성
정지 신호 - 엄지손가락 들기	높음	0.9238
정지 신호 - 왼쪽으로 밀기	중간	0.9370
정지 신호 - 제스처 없음	낮음	0.9513

Ⅵ. 결 론

본 연구에서는 시공간 입력 데이터의 구간 선택이 손 제스처 인식을 위한 3D-CNN모델의 학습 정확도에 미치는 영향을 분석하였다. 동적인 시공간 입력 데이터 구간에서 명확한 행동이 나타날 때 3D-CNN모델의 학습 정확도가 증가하고, 동적인 입력 데이터의 행동 변경 구간이 좁거나 너무 넓으면 학습 정확도가 떨어지는 것을 확인할 수 있었다. 또한, 학습 시 입력되는 동적인 시공간 데이터의 클래스 간 유사도가 높을수록 모델의 정확도가 낮아진다는 것을 알 수 있었다.

따라서 손 제스처 인식에 따른 동적인 시공간 데이터의 학습을 위한 연산 성능에 한계가 있거나 더 깊은 학습 모델을 설계할 수 없을 경우, 3D-CNN의 시공간 입력 데이터의 입력 구간을 변경하거나, 시공간 데이터의 클래스 별 상관관계 분석을 통해 학습 정확성을 높일 수 있음을 보여주었다.

다만, 동적인 시공간 데이터가 [멈춤(Stops), 동작(Actions), 멈춤(Shift-Stops)] 구간의 행동 주기를 따른다는 가정으로 데이터를 선정하였기 때문에 실험의 결과는 동적인 손 제스처 학습을 위한 모든 데이터가 [멈춤, 동작, 멈춤] 주기를 따른다는 한계가 있을 수 있다. 또한, 동적인 입력 데이터의 해상도가 높을수록 교차 상관 측정에는 많은 연산을 필요로 하고 학습 정확성은 동적인 시공간 입력 데이터 간의 유사성뿐만 아니라 입력 데이터의 절대적인 품질에 따라 달라지기 때문에 학습 시간에 영향을 미칠 수 있다는 한계도 있다.

향후에는 이러한 한계를 극복하기 위해, 동적인 시공간 데이터의 행동 변경 간격을 능동적으로 감지하여 입력 프레임 구간을 선택하는 기술과 클래스 간 유사도 측정을 위한 교차 상관 분석을 수행하면서 학습 시간을 최적화할 수 있는 동적인 입력 데이터의 해상도를 조절하는 방법을 연구하고자 한다.

References

[1] K. Alex, S. Ilya and E. T. Geoffrey, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in neural information processing systems, Communications of the ACM, vol. 60, no. 6, pp: 84-90, May 2017.
DOI: https://doi.org/10.1145/3065386

[2] S. G. Choi, and W. Xu, "A Study on Person Re-Identification System using Enhanced RNN," The journal of the institute of internet, broadcasting and communication (JIIBC), v.17 no.2, pp. 15-23, Apr. 2017.
DOI: https://doi.org/10.7236/JIIBC.2017.17.2.15

[3] T. Du, B. Lyubomir, F. Rob, T. Lorenzo and P. Manohar, "Learning Spatiotemporal Features with 3D Convolutional Networks," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp: 4489-4497, Oct 2015.
DOI: https://doi.org/10.1109/iccv.2015.510

[4] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
DOI: https://doi.org/10.3115/v1/d14-1179

[5] Hochreiter and Schmidhuber, "LONG SHORT -TERM MEMORY," 1997.
DOI: https://doi.org/10.1162/neco.1997.9.8.1735

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 2014.
DOI: https://doi.org/10.1109/cvpr.2014.223

[7] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017.
DOI: https://doi.org/10.1109/cvpr.2017.502

[8] K. Yang, R. Li, P. Qiao, Q. Wang, D. Li, and Y. Dou, "Temporal Pyramid Relation Network for Video-Based Gesture Recognition," 2018 25th IEEE International Conference on Image Processing (ICIP), Oct. 2018.
DOI: https://doi.org/10.1109/icip.2018.8451700

[9] T. Kim, D. Lim "A Study on Image Labeling Technique for Deep-Learning-Based Multinational Tanks Detection Model", The Journal of The Institute of Internet, Broadcasting and Communication, Vol.14, No.4, pp.58-63, 2022
DOI: http://dx.doi.org/10.7236/IJIBC.2022.14.4.58

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016.
DOI: https://doi.org/10.1109/cvpr.2016.213

[11] Twenty BN, "Jester dataset: a hand gesture dataset," https://www.twentybn.com/datasets/jester, 2017.

[12] J. David, "Correlation and Convolution", Class Notes for CMSC 426, 2005.

[13] Y. S. Kwon, S. Y. Hwang, D. J. Shin, J. J. Kim, "A Study on Application Method of Contour Image Learning to improve the Accuracy of CNN by Data", The Journal of The Institute of Internet, Broadcasting

and Communication (IIBC), Vol. 22, No. 4, pp.171-176, Aug. 31, 2022.
DOI: https://doi.org/10.7236/JIIBC.2022.22.4.171

[14] Y. Cho1, J. Kim, "A Study on The Classification of Target-objects with The Deep-learning Model in The Vision-images", Journal of the Korea Academia-Industrial Cooperation Society, Vol. 22, No. 2 pp. 20-25, 2021.
DOI: https://doi.org/10.5762/KAIS.2021.22.2.20

[15] G. W. Lee, J. H. Maeng, and S. Song, "Content based Image Retrieval Method that Combining CNNBased Image Features and Object Recognition Information", Journal of KIIT. Vol. 20, No. 5, pp. 31-37, May 31, 2022.
DOI: https://dx.doi.org/10.14801/jkiit.2022.20.5.31

저 자 소 개

정 영 지(정회원)

- 1995년 ~ 원광대학교 컴퓨터•소프트웨어공학과 교수
- 1993년 ~ 1995년 : 한국전자통신연구원
- 1987년 ~ 1993년 : 삼성 종합기술원
- 연세대학교 전기공학과 공학박사
- 관심분야 : 컴퓨터 네트워크, 인공지능 영상처리, 모바일 컴퓨팅