JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 1, pp. 55-61, January 2020
https://doi.org/10.9708/jksci.2020.25.01.055

A New Residual Attention Network based on Attention Models for Human Action Recognition in Video

Jee-Hyun Kim*, Young-Im Cho**

*Professor, Dept. of Software Engineering, Seoil University, Seoul, Korea
**Professor, Dept. of Computer Engineering, Gachon University, Sungnamsi, Korea

[Abstract]

With the development of deep learning technology and advances in computing power, video-based research is now gaining more and more attention. Video data contains a large amount of temporal and spatial information, which is the biggest difference compared with image data. It has a larger amount of data. It has attracted intense attention in computer vision. Among them, motion recognition is one of the research focuses. However, the action recognition of human in the video is extremely complex and challenging subject. Based on many research in human beings, we have found that artificial intelligence-like attention mechanisms are an efficient model for cognition. This efficient model is ideal for processing image information and complex continuous video information. We introduce this attention mechanism into video action recognition, paying attention to human actions in video and effectively improving recognition efficiency. In this paper, we propose a new 3D residual attention network using convolutional neural network based on two attention models to identify human action behavior in the video. An evaluation result of our model showed up to 90.7% accuracy.

▸ **Key words**: Deep Learning, Convolution Neural Network, Attention Mechanism, Video Processing, Action Recognition

[요 약]

딥 러닝 기술의 발전과 컴퓨팅 파워 등의 개선으로 인해 비디오 기반 연구는 최근 많은 관심을 얻고 있다. 비디오 데이터가 이미지 데이터와 비교하여 가장 큰 차이는 비디오 데이터에는 많은 양의 시간적, 공간적 정보가 포함되어 있다는 점이다. 이처럼 비디오에 포함된 많은 양의 데이터로 인해 컴퓨터 비전 연구에 있어서 행동 인식은 중요한 연구 과제 중 하나이지만, 비디오와 같이 움직임이 있는 환경에서 인간의 행동 인식은 매우 복잡하고 도전적인 과제이다. 인간에 대한 여러 연구를 바탕으로 인공지능에서는 인간과 유사한 주의(attention)메커니즘이 효율적인 인식 모델이라는 것을 알게 되었다. 이 효율적인 모델은 이미지 정보와 복잡한 연속 비디오 정보를 처리하는 데 이상적이다. 본 논문에서는 이러한 연구배경을 기반으로, 비디오에서 인간의 행동을 효율적으로 인식하기 위해 먼저 인간의 행동에 주목한 후 비디오 행동 인식에 주의메커니즘을 도입하고자 한다. 논문의 주요내용은 두 가지 주의 메카니즘을 기반으로 컨볼루션 신경망을 이용한 새로운 3D 잔류 주의 네트워크를 제안함으로써 비디오에서 인간의 행동을 식별하고자 한다. 제안 모델의 평가 결과 최대 90.7%정도의 정확도를 보였다.

▸ **주제어**: 딥 러닝, 컨볼루션 신경망, 주의 메커니즘, 비디오 프로세싱, 행동 인식

• First Author: Jee-Hyun Kim, Corresponding Author: Young-Im Cho
*Jee-Hyun Kim (jhkim@seoil.ac.kr), Dept. of Software Engineering, Seoil University
**Young-Im Cho (yicho@gachon.ac.kr), Dept. of Computer Engineering, Gachon University
• Received: 2019. 11. 18, Revised: 2020. 01. 20, Accepted: 2020. 01. 20.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 딥 러닝기술은 물체 감지, 이미지 분류, 이미지 분할, 얼굴 인식 및 자율 주행 등의 여러 분야에서 탁월한 성능을 보여주고 있다. 인공지능에서 중요한 비디오 이미지 처리기술은 과거의 수동 마크 처리 방법에서 시작하여 최근의 딥 러닝 에 이르기까지 다양한 처리에 적용되었다. 실제로는 비디오에 많은 이미지 데이터가 저장되어 있기 때문에, 비디오에서 이미지를 다루는 것에 관한 연구는 컴퓨터 비전의 중요한 연구 테마가 되고 있다. 최근에, CNN (Convolutional Neural Network)은 스틸 이미지 프로세싱(still image processing)에 매우 성공적인 연구결과를 보여주었다[1]. 이 결과로부터 CNN은 동적 이미지와 비디오 인식 시스템을 처리하기 위해 사용하기 시작하였다.

인간 행동 인식(human action recognition)은 컴퓨터 비전 분야에서 매우 중요한 테마이며, 인간의 행동 탐지 및 비디오 감시 분야에서 폭넓게 적용되고 있다. 단순한 사진 인식과 달리 비디오 스트림에서 동작을 인식하는 것은 매우 어렵고 복잡하다. 그 이유는 다음과 같다.

첫째, 비디오 기반의 인간 행동 인식은 복잡한 행동 배경을 가지며, 모션의 복잡성과 가변성은 동작을 정확하게 인식하기 어렵게 하기 때문이다. 둘째, 이미지 데이터와 달리 비디오 데이터에는 시간 정보가 포함되어 있어서 비디오 분류에 매우 중요한 요소로 작용한다. 역동적이고 복잡한 환경에서는 사람의 위치 결정이 더 어려워지고 특정 시간 간격 동안 특정 동작을 쉽게 식별 할 수 없는 경우가 많이 발생한다. 비디오에서 사람의 행동을 식별하려면 비디오를 정확하게 분할하고 인간의 행동이 발생한 시간대를 찾은 다음 해당 동작을 식별해야 하는데, 이 과정이 매우 복잡하다.

본 논문의 목적은 비디오에서 인간의 행동을 정확하게 인식하는 것이므로, 인체를 정확하게 감지하고 복잡한 비디오 데이터에서 행동 특징을 추출하는 방법을 제안한다.

이를 위해 비디오에서 인간의 행동 기능을 더 잘 식별하기 위해 주의 메커니즘(attention mechanism)을 도입하고자 한다. 주의 메커니즘은 인간의 비전 연구에서 비롯된 것으로, 인지 과학에서 정보 처리의 병목 현상을 해결하기 위한 것이다. 즉, 인간은 정보처리를 위해 다른 가시적 정보를 무시하고 필요한 정보에 선택적으로 집중하는 특징이 있다. 이 주의 메커니즘은 복잡한 정보를 갖는 비디오 데이터를 처리 할 때 매우 효과적인 알고리즘이다.

본 논문에서는 주의 메카니즘을 기반으로 컨볼루션 알고리즘을 이용한 새로운 3D 잔류 주의 네트워크(3D Residual Attention Network)를 제안하여 비디오에서 인간의 행동을 식별하고자 한다. 3D 컨볼루션 네트워크에는 공간 데이터를 처리하는 기능이 있어서 비디오 데이터를 보다 효과적으로 처리가능하다. 또한 잔류 네트워크에 주의 메카니즘을 추가하여 인간의 행동 기능에 집중하게 함으로써 인간의 행동을 효과적으로 인식할 수 있다.

본 논문의 구성은 다음과 같다. II장에서는 인간의 행동 인식 연구에 대한 기본배경에 대해 설명하고, III장에서는 본 논문이 제안하는 비디오에서 인간의 행동인식을 위한 3D 잔류 주의 네트워크 모델에 대해 설명하며, IV장에서 본 논문에서 개발한 알고리즘의 성능평가를 위한 실험결과를 제시하고 마지막 V장에서 결론을 맺고자 한다.

II. Related Work

인간 행동 인식은 컴퓨터 비전 분야에서 중요한 주제이며, 오랫동안 연구해온 인공지능의 한 분야이다. 단순한 이미지 인식과 달리, 사람의 행동 인식은 조명, 배경 등과 같은 많은 요소에 의해 방해를 받기 때문에 복잡한 기술이 필요하다.

인간의 행동 인식에 대한 기존의 전통적 연구방법은 다음과 같다. 한 데이터 세트 내에 있는 특정 동작 인식을 하기 위해, 일반적으로 STIP (Space-Time Interest Points), 그라디언트 히스토그램 [2] 및 광학 흐름 히스토그램, 3D 그라디언트 히스토그램 [3], 최첨단 수 공법의 고밀도 궤적 (iDT) [4] 등의 방법을 적용한다. 이러한 고밀도 궤적을 따라 풍부한 디스크립터를 풀링하여 행동 기능을 명시적으로 나타낸 후 카메라 움직임을 보정하는 기술을 적용한다. 이후에는 인코딩 방법에 의해 디스크립터가 비디오 레벨을 표현하는 기술을 사용함으로써 인간의 행동을 인식하여 왔다[5].

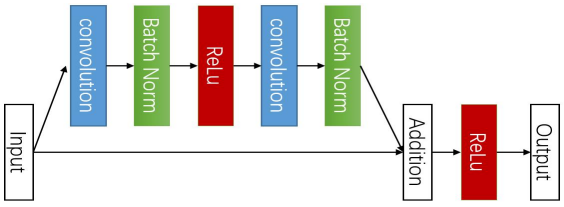

Fig. 1. Residual Network

최근에는 인공지능 기술의 발전으로 딥 러닝 기술을 인간 행동 인식 작업에도 적용하는 연구가 이루어지고 있다. Fig. 1은 잔류 네트워크의 일반적인 모델을 나타낸 것이다. 딥 러닝 알고리즘인 컨볼루션 신경망(Convolutional

Neural Network)은 입력 데이터에 대한 컨볼루션 연산을 통해 레이어별로 물체를 추출한 다음 이미지를 분류한다. 이러한 방법은 이미지 인식 분야에서 탁월한 결과를 보여주고 있다.

컨볼루션 신경망을 적용하여 우수한 성능을 보인 사례는 다음과 같다. 2012 AlexNet 네트워크[6]는 ImageNet 데이터 세트에서 상위 5개의 오류율을 16.4%로 감소시켰다. 2015 Inception v2 네트워크[7]는 배치 정규화(Batch Normalization) 방법을 제안하였으며, 2017 SeNet 네트워크[8]는 ILSVRC 대회에서 여러번 우승한 바 있다.

그러나, 비디오 인간 행동 인식 문제의 경우에는 비디오 프레임과 프레임 사이의 시간 상관관계로 인해 추출된 RGB 데이터 분류를 위해 컨볼루션 신경망으로 분류하는 것만으로는 좋은 결과를 얻지 못한다. 비디오 데이터 및 RGB 데이터가 매우 조밀하게 광학적으로 흐름을 형성하고 있을 때, 학습을 위해 이러한 정보들이 CNN으로 전송되고, CNN 신경망이 공간-시간(spatial-temporal) 정보를 잘 처리한 다음, 듀얼 스트림 네트워크(Dual Stream Network)에 의해 얻어진 결과들이 병합된다[9].

또 다른 연구로는 커널이 처리를 위해 연결된 후, 3D 컨볼루션 신경망이 인간의 행동 인식을 위한 정보를 훈련하고 추출하는 데 사용되기도 한다[10].

III. Method

1. Proposed Residual Attention Network

본 논문에서 제안하는 방법은 Fig. 1의 잔류 네트워크[1]를 기반으로 Fig. 2와 같이 구성한다.

기본 ResNets 블록은 2개의 컨볼루션 레이어로 구성되며 각 컨볼루션 레이어에는 배치 정규화와 ReLU가 있다. 바로 가기 패스는 블록의 상단을 블록의 마지막 ReLU 직전의 레이어에 연결하는 역할을 한다. 3D 컨볼루션 및 3D 풀링을 추가하여 3D 잔류 네트워크를 형성하여 비디오 공간-시간 데이터에 더 잘 적응하도록 한다. 이를 바탕으로 여러 주의 모듈(attention module)들을 쌓아서 잔류 주의 네트워크(residual attention network)를 구축하고자 한다. 각각의 주의 모듈에는 시간 및 공간 데이터를 개별적으로 처리하기 위한 두 가지 주의 모듈 즉, 채널 주의(channel attention)와 공간 주의(spatial sttention))를 포함한다는 점이 기존 3D 잔류 네트워크와 다른 주요 특징이다.

Fig. 2는 본 논문에서 제안하는 모델로서, 3D 잔류 네트워크 잔류 블록(Residual Network Residual Block)에서 잔류 주의 모듈(residual attention module)을 형성하기 위해 채널 주의 및 공간 주의 모델을 추가한 그림이다. 3D 컨볼루션 신경망을 적용한 후 채널은 피쳐 맵(feature map)에 주의를 기울인 다음, 공간 주의 모듈이 공간주의 기능 맵을 계산하는 방식이며 이점이 다른 잔류 네트워크와 비교해 차별화된 장점이다.

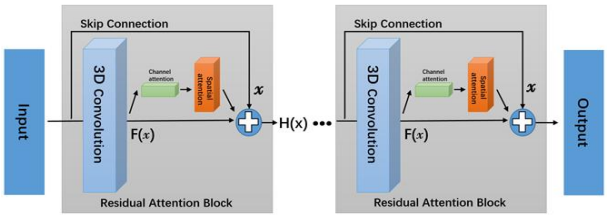

Fig. 2. Our Proposed 3D Residual Network Residual Block

Fig. 2에서 lth 레이어에 있는 합성 함수(composite function) Hl 은 각 피쳐 $|x_i|\begin{cases} l-1 \\ i=0 \end{cases}$ 3D 피쳐 맵을 입력 값으로 받는다. lth 레이어에 있는 출력 피쳐 맵인 Hl 은 다음 식(1)과 같다.

$$xl = \mathrm{H}l([x0,x1,x2...xl-1]) \quad (1)$$

여기서 $[x0,x1,x2...xl-1]$는 피쳐 맵들이 연결되어 있음을 나타낸 식이다. 각 피쳐 $|x_i|$의 공간 크기는 모두 동일하며, Hl(∗)는 BN-ReLU-3D 컨볼루션 연산에서 합성 함수를 의미한다.

2. Attention Models

2.1 Channel Attention Model

본 논문에서 사용하는 두 가지 주의 모델은 다음 Fig. 3과 같다. 먼저 채널 주의를 설명하면 다음과 같다.

Fig. 2에서 3D 컨벌루션 적용 후, 최대 풀링(max pooling) 및 평균 풀링(average pooling)의 조합에 의해 피쳐 맵의 크기가 압축되고, 다층 퍼셉트론의 히든 레이어가 파라미터를 감소시키기 위해 추가적으로 사용되며, 최종적으로 채널 주의 피쳐가 형성된다. 채널 주의는 이미지 기능에 주의하고 "무엇(what)"에 중점을 둔다.

요약하면, 채널 주의는 다음 식(2)와 같이 계산된다. 여기서 σ는 시그모이드 함수를 의미하고, 최종적으로 형성된 채널 주의 피쳐는 F_c로 표현하며, $F_{avg}{}^c$와 $F_{max}{}^c$는 각각 평균

풀링과 최대 풀링을 적용한 채널 주의 피쳐를, MLP(Multi Layer Perceptron)는 다층 퍼셉트론을 의미한다.

$$F_c = \sigma(MLP(F^c_{avg})) + (MLP(F^c_{\max})) \quad (2)$$

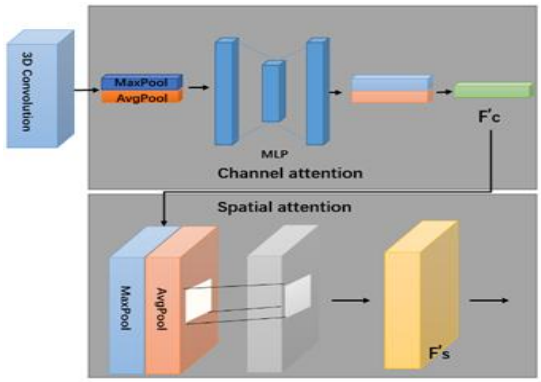

Fig. 3. Two Attention Models

2.2 Spatial Attention Model

본 논문에서 사용하는 두 가지 주의 모델 중 공간 주의를 설명하면 다음과 같다.

공간주의는 주로 기능 정보의 공간 위치에 주의하며 더 나아가 채널 주의에 의해 생성된 기능에 중점을 둔다. 여기서는 평균 풀링 및 최대 풀링 레이어가 사용되며, 생성된 피쳐는 병합되어 컨볼루션 레이어의 컨볼루션 알고리즘 적용 후 공간 주의 피쳐 맵을 형성한다.

요약하면, 공간 주의는 다음 식(3)과 같이 계산된다. 여기서 σ는 시그모이드 함수를 의미하고, 최종적으로 형성된 공간 주의 피쳐는 Fc로 표현하며, Favgc와 Fmaxc는 각각 평균 풀링과 최대 풀링을 적용한 공간 주의 피쳐를 의미한다. 각□X□은 컨볼루션 신경망에서 필터 크기를 7×7로 했을 때를 의미한다.

$$F_s = \sigma(f^{7\times 7}(F^s_{avg}; F^s_{\max})) \quad (3)$$

IV. Experiments

1. Datasets

본 논문에서는 개발한 알고리즘의 성능평가를 위해 잘 알려진 두 가지 데이터 세트 UCF-101[11]과 HMDB-51[12]을 사용하였다.

1.1 UCF-101

UCF-101은 YouTube에서 수집한 실제적인 액션 비디오 데이터베이스를 포함하고 있는데, 101개의 서로 다른 카테고리에서 13,320개의 짧은 크기의 비디오를 포함하고 있다. 각 비디오의 평균 길이는 약 7초이며, 동작 유형에는 인간과 물체 간의 상호 작용, 인간과의 상호 작용, 악기 연주, 스포츠 및 신체 운동 등 만을 포함하는 벤치마킹 데이터 세트이다.

Fig. 4. UCF-101 Data Set

1.2 HMDB-51

HMDB-51 데이터 세트는 대부분의 비디오 데이터를 영화 또는 공용 데이터베이스 및 YouTube와 같은 온라인 비디오 라이브러리에서 가져와서 구성한다. 이 데이터 세트에는 51개 이상의 카테고리가 있으며 각 카테고리 별로 약 101개 이상의 샘플을 포함하여 총 6,849개의 비디오 샘플이 포함되어 있는 벤치마킹 데이터 세트이다.

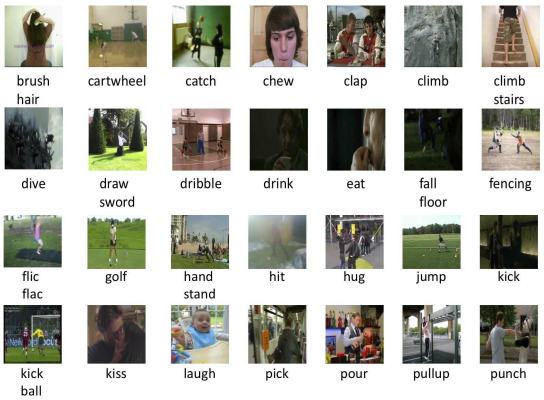

Fig. 5. HMDB-51 Data Set

2. Training Process

시뮬레이션을 위해 본 논문에서는 먼저, 데이터를 학습 세트(training set)와 테스트 세트(test set)로 나누어서 진행하였다. 즉, 위의 UCF-101과 HMDB-51 두 개의 데이터베이스로부터 얻은 데이터 세트 중에서 훈련 세트로 약 80%, 테스트 세트로 약 20%를 사용하여 실험을 진행하였다.

효율성 평가를 위해 GPU 로드에 대한 계산량의 한계점을 고려하여, 비디오 데이터를 초당 16프레임 (16fps)으로 분할하고, 중앙 자르기를 사용하여 비디오 이미지 크기를 각각 224 × 224 크기로 잘라서 사용하였다. 네트워크는 확률적 경사 하강 (stochastic gradient descent, SGD)[13] 알고리즘을 사용하여 학습시켰다. 실험에서 초기 학습 속도를 0.1로 설정하고 점차 줄여 나가는 방식으로 하였으며, 0.9 Nesterov 모멘텀 값[14]을 사용하였고, 모든 가중치에 대한 감소없이 10-4 가중치 감소를 사용하였다. 네트워크 성장 속도는 12에서 24로 설정하였고, 배치 크기는 8에서 10으로 설정하였으며, 100번의 컨볼루션 신경망 에포크(epoch) 반복으로 학습시켰다.

3. Performance Evaluation

실험결과, 본 논문에서 제안한 모델은 다른 모델에 비해 높은 정확도를 가지고 있음을 알 수 있었다. 3D 모델에서 두 가지 주의 모델을 추가 한 3D 잔류 네트워크가 기존의 3D 잔류 네트워크보다 더 효율적이라는 것을 Table 1 및 Table 2의 데이터에서 알 수 있다. 이는 유사한 방식의 논문[15]보다 두 가지 주의 모델의 계산방식을 단순화 시켰기 때문이기도 하다.

실험결과는 Table 1, Table 2에 제시하였다. Table 1에서는 UCF-101 데이터 세트를 세 가지 그룹으로 나눈 것 중 하나의 그룹(split)을 활용하여, 본 논문에서 제안한 방법과 기존의 다른 3D 잔류 네트워크와의 성능비교를 나타낸 것으로, 제안한 모델이 다른 방법들에 비해 정확도가 높게 나타났다. 제안한 모델은 사용한 데이터 세트에 대해 72.8%의 정확도를 보여주었다.

Table 1. Exploration of our model and other 3D ConvNets on the UCF-101 dataset (split1)

Method	Accuracy(%)
ResNet3D-50[1]	59.2
DenseNet3D[10]	68.5
Our model	72.8

Fig. 6은 오리지널 비디오 이미지와 이 비디오 이미지에 제안한 모델을 적용한 결과를 보여준 것으로, 아래 그림에서 파란색 선이 실제 관심영역이며 빨간색 선이 공간 주의 모델 적용 후의 모습으로 상당히 일치함을 알 수 있다.

Fig. 6. Original image(top) vs. Our mode image(bottom)

본 논문에서 제안한 방법은 시공간 정보 캡처 용량을 크게 향상시킬 수 있는데, 이유는 Fig. 6에서와 같이 전체 비디오 그림에서 주의 모델을 사용하여 캡쳐하기 때문이다. 그러나 비디오 마다 주의 모델이 다르므로 정확한 캡쳐용량은 예측하기 어려우나 비디오 이미지를 224 × 224 크기로 잘라서 사용하므로 적은 크기도로 정확도를 다른 네트워크에 비해 72.7%까지 향상시켜 주었다. 따라서 좀더 많은 비디오 이미지를 동시에 캡쳐하여 사용할 수 있어서 성능이 향상됨을 알 수 있다.

Table 2는 UCF-101 데이터 세트와 HMDB-51 데이터 세트에서 제공하는 세 가지 그룹(split)을 모두 활용하여 얻은 결과를 나타낸 것이다. 본 논문에서 제안한 방법과 기존의 다른 3D 잔류 네트워크와의 성능비교결과를 보면, 제안한 모델이 다른 방법들에 비해 정확도가 높게 나타났음을 알 수 있다. 제안한 모델은 두 가지 데이터 세트에 대해 각각 90.7%와 63.1%의 정확도를 보여주었다. Fig. 7은 이것을 그래프로 나타낸 것이다. 주의 모델에 의해 캡쳐된 정보로도 매우 높은 정확도를 보여주고 있다.

Table 2. Accuracy (%) performance comparison of our model with other methods over all three splits of UCF-101 and HMDB-51

Method	UCF-101(%)	HMDB-51(%)
C3D[16]	82.3	56.8
Conv. Fusion[17]	82.6	56.8
Two Stream[18]	88.6	-
ResNet3D[1]	86.1	55.6
DenseNet3D[10]	88.9	57.8
Our model	**90.7**	**63.1**

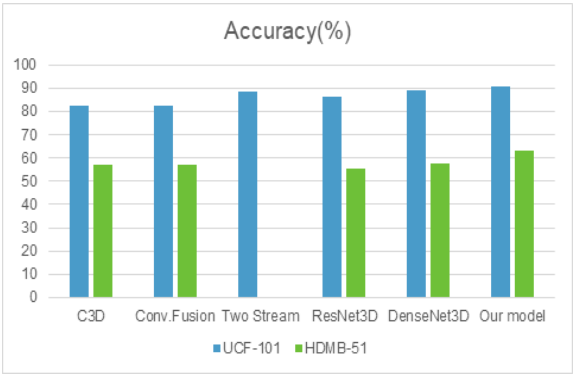

Fig. 7. Accuracy (%) performance comparison

V. Conclusion

본 논문에서는 주의 모델에 기반한 새로운 3D 잔류 네트워크를 제안하였다. 제안한 네트워크의 장점은 시공간 정보 캡처 용량을 크게 향상시킬 수 있다는 것이다. 여러 UCF-101 및 HMDB-51 데이터 세트를 활용하여 실험한 결과, 본 논문에서 제안한 3D 잔류 네트워크의 성능이 기존의 3D 잔류 네트워크 및 기타 네트워크의 성능보다 우수함을 알 수 있었는데 이유는 두 가지 주의 모델(채널 주의 및 공간 주의)을 사용하고 있기 때문이다.

향후 연구로는 광학적 흐름 맵(optical flow map) 기반에서 피쳐 캡쳐를 효율적으로 하는 방법과 컴퓨팅 오버헤드 감소를 위한 연구를 계속 진행하고자 한다.

ACKNOWLEDGEMENT

The present research has been conducted by the Research Grant of Seoil University in 2019.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] N. Dalal, B. Triggs, C. Schmid, “Human detection using oriented histograms of flow and appearance,” In ECCV, 2006.

[3] A. Klaser, M. Marszałek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” In BMVC, 2008.

[4] H. Wang and C. Schmid, “Action recognition with improved trajectories,” In P with improved trajectories. In Proc. ICCV, 2013.

[5] H. Wang and C. Schmid, “Action recognition with improved trajectories,” In ICCV, 2013.

[6] A. Krizhevsky, I. Sutskever , G. Hinton, “Elmagenet classification with deep convolutional neural networks[C],” Advances in Neural Information Processing Systems , 2012.

[7] S. IOFFE , C. SZEGEDY, “Batch normalization: accelerating deep network training by reducing internal covariate shift[C],” Proceedings of the 32nd International Conference on Machine Learning , 2015.

[8] J. HU , L. SHEN , G. SUN, “Squeeze-and-excitation networks[J],” arXiv preprint arXiv : 1709.01507 , 2017.

[9] L. WANG , Y. XIONG , Z. WANG , et al., “Temporal segment networks : Towards good practices for deep action recognition[C],” European Conference on Computer Vision. Springer , Cham , 2016.

[10] S. JI , W. XU , M. YANG , et al., “3D convolutional neural networks for human action recognition[J],” IEEE Transactions on Pattern Analysis and Machine Intelligence , 2013.

[11] K. Soomro, A.R. Zamir, M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” 2012.

[12] H. Kuehne, H. Jhuang, R. Stiefelhagen, T. Serre, “ Hmdb51: a large video database for human motion recognition.,” High Perform. Comput. Sci. Eng. 12, pp. 571–582 2013.

[13] M. Zinkevich, M. Weimer, L. Li, A. Smola, “ Parallelized stochastic gradient descent,” In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, pp. 6–9 December 2010.

[14] I. Sutskever, J. Martens, G. Dahl, G. Hinton, “On the importance of initialization and momentum in deep learning,” In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, pp. 16–21 June 2013.

[15] Jiahui Cai, Jianguo Hu, “3D RANs: 3D Residual Attention Networks for action recognition,” The Visual Computer, 25, July 2019.

[16] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497, 2015.

[17] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” In Proc. CVPR, 2016.

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, “Convolutional two-stream network fusion for video action recognition,” In Proc. CVPR, 2016.

Authors

Jee-Hyun Kim received a Doctor of Computer Science and M.B.A degree in the Information Management from Dankook University, Korea, in 2004 and 1994, respectively. Her B.S degree is Mathematics from Ewha Womans University in 1978. She is a professor at Seoil University. She is interested in Web Engineering, Big data, Quality Management, Information Management etc.

Young-Im Cho received her B.S., M.Sc., and Ph.D from the Department of Computer Science, Korea University, Korea, in 1988, 1990 and 1994, respectively. She is a professor at Gachon University. Her research interest includes AI, Big data, information retrieval, smart city etc.