영상 데이터에서의 CRNN 기반 자세 인식*

임세민 O1 채병철 2 임수빈 2 박주영 3 오형철 3
[1] 고려대학교 병렬연산연구실
[2] 고려대학교 전자및정보공학과
[3] 고려대학교 과학기술대학
{jaewoong819, parkj, ohyeong}@korea.ac.kr

CRNN-based Human Posture Recognition in Video Data

Se-Min Lim O1 Byeong-Cheol Chae 2 Soo-Bin Lim 2 Jooyoung Park 3 Hyeong-Cheol Oh 3
[1] Parallel Computation Lab., Korea University
[2] Dept of Electronics and Information Eng., Korea University
[3] College of Science and Tech., Korea University

요 약

영상 데이터로부터 사람의 자세를 인식하는 시스템을 설계하였다. 설계한 시스템은 합성곱 신경망과 GRU 순환 신경망으로 이루어진 end-to-end 신경망으로 구성되었으며, 휴대폰 카메라로 수집된 자세나 동작을 연습하는 사람의 영상 데이터를 이용하여 학습되어, 연습하는 사람의 동작이 어떤 동작에 해당하는가를 추론하여 알려줌으로써, 연습에 도움을 제공할 수 있다. 국민 체조에서 사용되는 4가지 기본 동작들의 영상 데이터를 수집하여 실험해 본 결과, 개발된 시스템은 비교적 간편한 방법으로 다양한 활동의 훈련이나 재활을 도와주는 시스템의 기초를 제공할 것으로 기대된다.

1. 서 론

최근 사람의 움직임에 관한 정보를 분석하여 행동을 인식하는 기술들이 활발하게 연구되고 있는데, 본 논문에서는 그 중 숙련도 평가(Skill Assessment) 문제[1,2,3]를 연구하였다. 복잡한 동작의 경우에, 사람의 몸에 부착된 센서데이터를 이용하여 좋은 결과를 얻을 수 있음을 보이는 연구결과들이 발표되고 있으나[2,3], 본 논문에서는, 시계열 영상 인식에 최근 활발히 적용되는 CRNN(Convolutional Recurrent Neural Network)[4]을 사용하여 영상데이터로부터 체조연습을 보조하는 시스템을 제안하고 성능을 평가하였다.

제안하는 시스템은, 체조 동작을 휴대폰 카메라로 촬영한 영상 데이터를 사용하여 사용자가 취한 체조의 기본 동작이 시스템에 학습되어 있는 여러 동작 중 어느 동작에 해당하는 가를 사용자에게 알려준다.

입력 영상이 연속적인 이미지 프레임의 구성이라는 점을 고려하여, 합성곱 신경망(Convolutional Neural Network; CNN)과 순환 신경망(Recurrent Neural Network; RNN)이 연결된 하나의 end-to-end CRNN 형태의 신경망으로 설계하였다. 추론 성능과 저비용 측면을 고려하여, 합성곱 신경망으로는 Inception-v3 Pretrained 모델[5]을 사용하였고, 순환신경망으로는 GRU(Gated Recurrent Unit) RNN[6]을 사용하였다. 제안하는 시스템은 약 94%의 추론 정확도를 보였으며, 비교적 간편한 시스템을 사용하여 사람의 자세 인식 및 교정에 도움을 주는 것이 가능함을 보여주었다.

2. 제안하는 시스템의 설계

2.1 데이터 수집

그림 1과 같이, 휴대폰 카메라를 이용하여 각 2명씩의 초보자와 숙련자가 기본 4가지 체조동작을 수행한 모습을 영상으로 촬영하여 데이터 셋을 구성하였다. 한 동작 당 10번을 촬영하여, 총 160개의 영상 데이터를 수집하였으며, 이 중 70%는 학습(training)용으로, 30%는 테스트(test)용으로 사용하였다.

*본 연구는 한국연구재단의 중견과제 (과제 번호:2017R1E1A1A03070652)에서 지원하여 연구하였음.

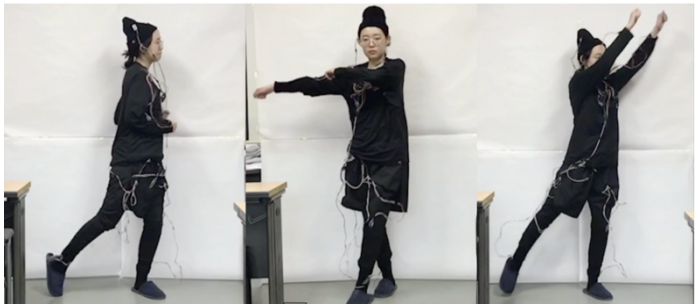

그림 1. 데이터 수집

2.2 이미지 프레임 추출

비디오 데이터는 모든 신경망의 학습 및 추론 데이터로 사용하기 매우 한정적이므로, 촬영한 영상 데이터를 이미지 프레임(Frame)으로 추출하였다. 1 fps(frame per second)로 jpg 형식의 프레임을 각 영상별로 추출하였으며, 시스템 설계의 간소화를 위해 다른 이미지 전처리(Preprocessing) 작업은 수행하지 않았다.

2.3 합성곱 신경망

합성곱 신경망은 이미지 인식분야에서 사용되는 가장 대표적인 신경망으로써, 이미지의 특징점(Feature)들을 찾아내는데 매우 특화되어 있으며, 다른 신경망들보다 해당 분야에서의 가장 높은 추론 성능을 보여주고 있다. 본 논문에서 추출한 이미지 프레임에서의 적절한 특징점들을 찾아내기 위해, 그림 2와 같은 합성곱 신경망을 설계하였다.

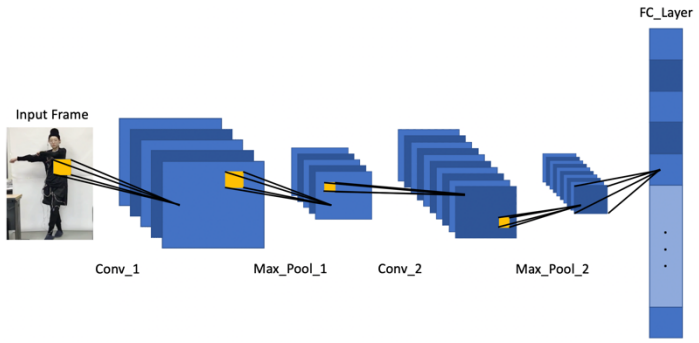

그림 2. 설계한 기본 합성곱 신경망

또한, 추가적으로 구글 (Google) 사의 Inception-v3 Pretrained 모델[5]을 사용하였다. Pretrained 모델을 사용하면, 이미 이미지 인식에서 매우 높은 추론 성능을 보여주고 있는 해당 모델의 파라미터(Parameter)들을 그대로 가져올 수 있으며, 이를 기반으로 새로운 데이터셋에 재학습(Retraining)만 수행해주면 된다는 장점이 있다. 본 논문에서는 Inception-v3 모델에, 수집한 체조 데이터를 재학습시켰으며, 설계한 기본적인 합성곱 신경망과 성능을 비교하였다.

2.4 순환 신경망

영상 데이터는 시간의 연속성을 가지고 있는 이미지 프레임 기반 시계열 데이터이다. 순환신경망은 시간의 연속성으로 만들어진 이전 및 이후 데이터 사이의 연관성 추출에 뛰어난 성능을 가지고 있다. 효율적인 추론 성능을 도출하기 위해, 본 논문에서는 합성곱 신경망 이외에 추가적으로 순환신경망을 사용하였다.

본 논문은 그림 3과 그림4에 보인, LSTM(Long Short-Term Memory)[3]와 GRU(Gated Recurrent Unit)[6]를 고려하였다. 두 신경망 모두 단층(1-Stacked) 및 단방향(Unidirectional) 조건만을 설정하여 설계하였고, 추론 성능 및 메모리 사용량 측면에서 비교 분석하였다.

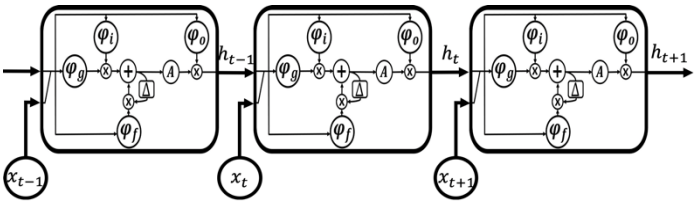

그림 3. LSTM[3]

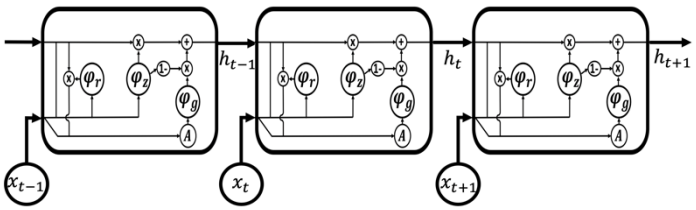

그림 4. GRU[6]

2.5 End-to-end CRNN 신경망

영상 데이터가 이미지 프레임이 연속적으로 연결되어 있는 형태라는 점을 고려하여, 이미지의 특징점들을 효율적으로 분석하는 합성곱 신경망에 time-series 이미지 데이터를 효율적으로 학습 및 추론하기 위한 순환 신경망을 end-to-end 방식으로 연결하여 하나의 CRNN[4]의 형태로 설계하였다.

그림 5은 설계의 순서를 보인 것이다. 영상 데이터를 이미지 프레임으로 추출하여, 프레임의 특징점들을 추출하는 합성곱 신경망을 사용한다. 그 다음, 추출한 특징점들을 인코딩(Encoding)하고, 해당 결과와 영상에 맞게 연속적인 이미지 프레임의 순서를 기록(Recoding)한다. 마지막으로 순환 신경망에서, 인코딩 되어 있는 데이터들을 디코딩(Decoding)하며 4가지 동작과 4명의 선수로 구성되어 있는 16개의 클래스(Class)에 맞게 레이블(Label)이 기록되어 있는 파일을 참조하여 지도

학습(Supervised Learning)을 진행한다.

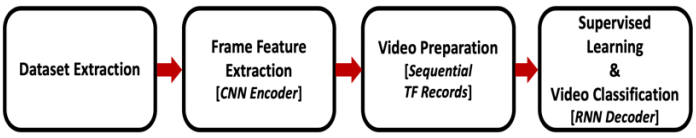

그림 5. 설계의 순서도

3. 결 과

본 논문의 시스템은, Intel i5-8300HU CPU와 NVIDIA GTX 1060 GPU 및 Ubuntu 18.04 환경 상에서, Tensorflow[7]를 Backend로, Keras를 Frontend로 사용하여 설계하였다.

표 1은 기본적인 합성곱 신경망을 인코더(encoder)로 사용한 경우에, 설계한 신경망의 총정확도(Accuracy)와 top-5 정확도(Top-5 Accuracy)를 보인 것이다. 3가지 타입의 신경망 모두 비교적 낮은 정확도를 보였으며, 이미지 프레임에서의 특징점 추출이 매우 중요한 과제임을 보이고 있다.

표 1. 기본 합성곱 신경망 기반 CRNN의 정확도

CRNN 신경망	정확도	Top-5 정확도
CNN + RNN	52.08%	50.00%
CNN + LSTM	60.42%	62.50%
CNN + GRU	62.50%	60.42%

표 2는 Inception-v3 Pretrained 모델[5]을 인코더로 사용하였을 때의 정확도를 보인 것이다. 3가지 신경망 모두 비교적 높은 정확도를 보였으며, 특히, LSTM[3]과 GRU[6] 셀(Cell)을 사용할 때, 90% 이상의 높은 정확도를 보였다. 표 1과 비교하여 보면, 정확한 특징점 추출을 위해 인코더의 선택도 중요하지만, 디코더(Decoder)의 사용에서 그라디언트 손실(Gadient Vanishing) 문제가 있는 기본적인 RNN 셀[8]보다는 LSTM이나 GRU 사용이 더 나은 선택임을 알 수 있다.

표 2. Inception-V3 모델 기반 CRNN의 정확도

End-to-end 신경망	정확도	Top-5 정확도
Inception-v3 + RNN	83.33%	85.42%
Inception-v3 + LSTM	93.75%	93.75%
Inception-v3 + GRU	93.75%	91.67%

표 3은 디코더의 타입 별로 학습 파라미터 수를 비교하고 있다. 기본적인 RNN[8]은 리소스 사용량이 매우 적지만, 정확도면에서 우수한 성능을 보여주지 못하고 있기 때문에, 추론 성능과 파라미터 사용량을 고려하였을 시, GRU가 디코더로서 적합함을 보였다.

표 3. 리소스 사용량 [개]

순환신경망	Simple RNN	LSTM	GRU
학습 파라미터 수	43,501	167,641	126,261

참고 문헌

[1] M. Kranz, et al., “The Mobile fitness coach: Towards individualized skill assessment using personalized mobile devices.” Pervasive and Mobile Computing. 9(2), 2013, pp.203-215.

[2] 임세민 외, “딥러닝을 이용한 탁구연습 보조시스템”, 한국정보과학회 KSC논문집, 2017년 12월, pp.960-962.

[3] S.-M. Lim, et al., “LSTM-guided coaching assistant for table tennis practice,” MDPI Sensors, 2018, 18(12), pp.4112-4126.

[4] B. Shi, et al., “An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” IEEE Trans. PAMI, 39(11), Nov. 2017, pp.2298-2304.

[5] C. Szegedy, et al., “Rethinking the inception architecture for computer vision,” IEEE Conf. CVPR, 2016, pp.2818-2826.

[6] J. Chung, et al., “Empirical evaluation of gated recurrent neural networks on sequence modeling,” ArXiv:1412.3555, 2014.

[7] M. Abadi, et al., “Tensorflow: A system for large-scale machine learning,” USENIX Symp. on OSDI 16, 2016, pp.265-283.

[8] T. Mikolov, et al., “Recurrent neural network based language model,” In INTERSPEECH-2010, pp.1045-1048.