Journal of Korea Robotics Society (2018) 13(1):045-054
https://doi.org/10.7746/jkros.2018.13.1.045 ISSN: 1975-6291 / eISSN: 2287-3961

열화상 카메라를 이용한 3D 컨볼루션 신경망 기반 낙상 인식

3D Convolutional Neural Networks based Fall Detection with Thermal Camera

김 대 언[1] · 전 봉 규[2] · 권 동 수†
Dae-Eon Kim[1], BongKyu Jeon[2], Dong-Soo Kwon†

Abstract: This paper presents a vision-based fall detection system to automatically monitor and detect people's fall accidents, particularly those of elderly people or patients. For video analysis, the system should be able to extract both spatial and temporal features so that the model captures appearance and motion information simultaneously. Our approach is based on 3-dimensional convolutional neural networks, which can learn spatiotemporal features. In addition, we adopts a thermal camera in order to handle several issues regarding usability, day and night surveillance and privacy concerns. We design a pan-tilt camera with two actuators to extend the range of view. Performance is evaluated on our thermal dataset: *TCL Fall Detection Dataset*. The proposed model achieves 90.2% average clip accuracy which is better than other approaches.

Keywords: 3D Convolutional neural networks, Fall detection, Thermal images

1. 서 론

의료 기술의 발달과 생활 수준의 향상으로 고령화 사회에 진입하면서 건강 문제가 중요시 되고 의료 서비스에 대한 국민적 관심이 높아지고 있다. 그 중에서도 낙상 인식은 헬스 케어, 무인 감시, 인체 행동 인식 등의 관련 분야와 함께 다양한 연구가 활발히 진행되고 있다. 특히 의료기관에서 입원환자의 낙상은 의료서비스 향상을 위해 해결해야 할 가장 중요한 문제로 대두되고 있다. 낙상을 당한 환자는 골절, 외상 후 스트레스 장애, 사망 등 환자 안전에 직결된 문제를 야기시키고 결과적으로 환자 및 보호자의 치료 시간과 비용을 가중시키고, 질병치료에도 부정적인 영향을 미치게 되며, 의료과실 소송과 같은 법적인 분쟁도 일으키게 된다[1].

세계 보건 기구(World Health Organization)의 발표[2]에 따르면, 매년 약 28-35 %의 65세 이상 환자들이 낙상을 당하고 있으며 70세 이상의 환자들의 경우 낙상을 경험한 비율이 약 32-42 %에 달하는 것으로 조사되었다. 또한 보호자와 함께 생활하는 환자에 비해 요양원이나 시설, 병원에서 보호자 없이 혼자 생활하는 환자가 더욱 많은 낙상을 경험하는 것으로 알려졌다. 이와 같이 낙상 발생의 빈번함은 환자의 나이와 노쇠함의 정도에 따라 증가한다. 또한, 65세 이상 노인을 대상으로 하는 낙상 역학 조사[3]에 따르면 심혈관 질환, 암, 뇌졸중, 폐기관 장애에 이어 낙상이 노인의 주요 사망 요인으로 조사되었다. 따라서 홀로 생활하는 노인을 대상으로 하는 낙상 모니터링 시스템이 시급히 도입되어야 한다는 것을 알 수 있다.

낙상은 발생 후 최대한 빠른 시기에 발견하여 의료 조치를 하는 것이 매우 중요하며 조기에 발견하여 치료하지 못 할 경우 부상의 정도나 사망률이 급격히 증가하는 것으로 알려져 있다[4]. 서울 소재 500병상 규모의 종합병원에 입원한 환자 92명을 대상으로 한 조사[5]에 따르면, 낙상은 입원 후 1-5일 사이가 45명(48.9 %)으로 가장 많았으며, 그 중에서도 19명(20.7 %)으로 입원 후 2일째가 가장 많았다. 낙상이 일어난 시간대는 새벽(00:00-06:00)이 43명(46.7 %)으로 가장 많았고, 오전(06:00-12:00) 22명(23.9 %), 오후(12:00-19:00) 10명(10.8 %),

Received : Feb. 3. 2018; Revised : Feb. 19. 2018; Accepted : Feb. 19. 2018
※ This work was supported by the Industry Strategic Technology Development Program (No. 10052358, Development of Robot Systems for Total Nursing Service) funded by the Ministry of Trade, Industry, and Energy, Korea.
1. Ujin Technology, Daejeon, Korea (kimde1205@gmail.com)
2. Robotics Program, KAIST(Korea Advanced Institute of Science and Technology, Daejeon, Korea (bk_jeon@kaist.ac.kr)
† Corresponding author: Department of Mechanical Engineering, KAIST(Korea Advanced Institute of Science and Technology, Daejeon, Korea (kwonds@kaist.ac.kr)

밤(19:00-24:00) 17명(18.5 %) 순으로 야간에 가장 많은 낙상 사고가 발생하였다. 하지만 야간에는 제한된 인력으로 인해 환자에 대한 의료진의 보호와 관심이 부족한 상황이기 때문에 낙상 발생이 가장 빈번하나 조기 발견과 조치에 취약한 문제점이 있다. 따라서 주야간의 제한없이 환자의 낙상 인식과 알림을 통해 의료진의 빠른 의료 조치를 가능케 하는 환자 낙상 모니터링 시스템이 필요하다.

본 연구에서는 비디오 영상 분석에 적합한 시공간적 특징 학습에 적합한 심층 신경망 구조를 찾는데 초점을 맞추고 있다. 이를 위해 기존의 컴퓨터 비전 분야에서 뛰어난 인식 성능을 보이고 있는 CNN 모델의 도메인을 시간 축에 대해 확장한 3D 컨볼루션 연산을 적용하였다. 3D 컨볼루션 신경망은 인접한 이미지들의 집합으로 이루어진 영상 볼륨을 입력으로 받아 단일 이미지 내의 공간적인 정보뿐만 아니라 연속적인 프레임 사이의 순차적인 관계에 대한 정보까지 함께 포함하는 시공간적(Spatio-temporal) 정보를 학습하였다. 또한 환자 영역 추적을 위한 팬틸트 제어를 통해 열화상 카메라의 시야각 한계 문제를 해소하여 보다 넓은 영역을 모니터링 할 수 있도록 하였다.

2. 선행 연구

환자의 낙상을 인식하는 시스템을 개발하기 위해 지금까지 다양한 연구가 진행되어 왔다. 기존의 낙상 인식 시스템을 사용된 센서를 기준으로 나눠보면 주변 장치, 신체부착형 장치, 모바일 기기 내장 센서 등의 비영상 기반 연구와 시각 센서를 이용한 영상 기반 연구로 분류할 수 있다. 본 장에서는 비영상 기반 및 영상 기반 낙상 인식 연구 각각의 접근법과 그 한계점에 대해 서술하고 낙상 인식의 연관 분야로서 인체 행동 인식 연구에 대해 간략히 소개한다.

2.1 비영상 기반

주변 장치 기반 시스템은 음향 센서, 바닥 압력 센서 등의 장치를 사전에 설치하여 센서 영역에서 바닥의 압력 변화, 충격음 발생 등을 기준으로 낙상 여부를 판단한다. Rimminen[6]은 바닥 압력 센서를 이용한 낙상 인식 시스템을 제안했다. 실험 바닥 영역의 크기와 모양, 압력 센서의 크기를 계산하여 추출된 특징들을 Bayesian filtering 기법과 두 가지 상태를 가지는 Markov chain model을 사용하여 낙상 여부를 판단하였다. 하지만 실험 결과, 피실험자가 무릎을 사용하며 낙상을 하는 경우에는 그 패턴이 서 있는 상태와 매우 비슷하여 20 %의 매우 낮은 정확도를 나타내었다. 일반적으로 주변 장치 기반 시스템은 실제 환자가 생활하는 공간뿐만 아니라 낙상 인식에 불필요한 영역까지 장치가 설치된 모든 곳에서의 데이터를 감지하기 때문에 오경보를 빈번히 발생시켜 전체 낙상 인식률이 낮다는 치명적인 문제점이 있다.

신체부착형 장치 기반 시스템은 가속도계, 자이로스코프, 압력 센서 등의 소형 전자 센서를 신체에 부착하여 환자의 움직임이나 위치 정보를 수집하며 이를 분석하여 낙상 여부를 판단한다. 가속도계, 자이로스코프가 탑재된 소형 센서 장치를 엉덩이, 발목, 무릎, 손목, 어깨, 팔꿈치 부위에 부착하여 피실험자의 움직임이나 위치 정보를 추정하고, 가속도, 속도 분포값을 이용하여 특징 벡터를 구성하고, 이를 SVM 등의 분류기를 통하여 낙상을 인식하는 알고리즘을 제안했다[7, 8].

하지만 신체부착형 장치 기반 시스템은 센서의 비용이 저렴하며 센서의 소형화가 가능하다는 장점이 있지만 사용자가 신체에 부착하므로 일상생활 움직임에 제약이 발생하고, 노인이나 전자 장치에 익숙하지 않은 사용자가 거부감을 가진다는 한계점을 가진다[9].

또한 최근 모바일 기기의 휴대성과 그 성능에 힘입어 스마트폰에 내장된 가속도, 자이로스코프 센서를 이용한 낙상 인식 연구가 제안되었다[10, 11]. 스마트폰 센서 기반 연구에서는 데이터의 빠른 처리와 배터리 절약을 위해 특정 임계값을 기준으로 낙상을 인식하는 단순한 알고리즘을 채택하고 있어 패턴의 복잡도 또는 다양성에 대한 요구에는 한계가 있다.

2.2 영상 기반

시각 센서 기반 시스템은 카메라를 이용하여 촬영된 영상을 처리하여 환자의 낙상을 인식한다. 카메라를 이용할 경우, 동시다발적으로 발생하는 여러 이벤트를 한번에 감지할 수 있고 비교적 높은 정확도를 가져 최근 무인 감시, 헬스 케어 분야와 함께 낙상 인식 분야에서 크게 각광 받고 있다. 기존의 영상 기반 연구는 열화상 카메라 또는 깊이 카메라를 사용하여 영상 내 사람의 외형 변화를 이용하여 낙상 여부를 판단하였다[12, 13, 14, 15]. 배경 제거 방법(Background subtraction method)을 통해 영상 내 대상의 바운딩 박스(Bounding box)를 찾아 대상의 높이와 너비 정보를 추출한다. 추출된 높이와 너비 비율을 매프레임마다 계산하여 임계값 이하로 감소할 경우 낙상으로 인식하였다. Vadivelu[16]는 55개의 상황으로 이루어진 열 영상 데이터를 수집하고 HOG, HOF 특징으로 추출하여 SVM을 이용한 이진 분류를 통해 낙상 여부를 판단하였다.

시각 센서 기반의 시스템은 GPU, CPU 클러스터링 등의 컴퓨팅 능력을 통한 높은 성능을 보이고 있지만 일반적으로 조

명 밝기, 가려짐, 카메라 시점, 초점 변화 등 환경 변화에 대해 취약하다. 또한, 환자의 사생활 침해 관련 문제는 실시간 낙상 인식을 위한 데이터 획득 및 성능 향상을 위해 반드시 해결해야 할 과제로 남아있다.

2.3 인체 행동 인식

낙상 인식과 밀접한 연관을 갖는 인체 행동 인식(Human action recognition) 연구에서는 비디오 영상에 대한 인식 성능 향상을 위한 연구가 진행되어 왔다. 직접 가공한 특징(Hand-crafted features) 기반 연구에서는 SIFT-3D[17], HOG-3D[18]와 iDT[19] 등의 특징을 통해 시공간 정보를 추출하고자 하였다. 또한 UCF-101, Sports-1M 등의 대용량 비디오 데이터 분석을 위한 방안으로서 3D 컨볼루션 신경망 모델이 제안되었다[20, 21]. 또한 이미지의 외형에 담긴 정보를 학습하는 경로와 프레임 사이의 관계에 의한 시간적 정보를 추출하는 경로를 개별적으로 구성하고 지연 결합(Late fusion)을 통해 최종 분류를 수행하는 Two-stream 네트워크가 사용되기도 하였다[22].

본 논문에서는 인식 성능 향상을 위해 3D 컨볼루션 신경망을 기반으로 시계열 영상에 내포된 시공간적 특징을 학습하고 열화상 카메라를 이용한 주야간 감시, 환자 사생활 문제를 해소하는 낙상 모니터링 시스템을 제안하였다.

3. 3D 컨볼루션 신경망

비디오 영상 분석을 위해서는 영상의 순차적인 흐름에 따른 시간적 정보를 공간적인 정보와 함께 담아내는 것이 매우 중요하다. 2D 컨볼루션 신경망은 수많은 영상 인식 분야에서 높은 성능을 강점으로 현재까지 많은 연구가 이뤄져 왔다. 하지만 컨볼루션 신경망은 단일 프레임 입력을 다루는 것으로 한정되어 있다. 따라서 다수의 인접한 프레임으로 구성된 비디오 영상에 내재된 모션 정보를 담을 수 없다. 이에 반해 3D 컨볼루션 신경망은 외형과 모션에 관한 정보를 동시에 특징으로 추출할 수 있으며 비디오 분석에서 2D 컨볼루션 신경망 구조를 능가하는 성능을 보인다.

최근 3D 컨볼루션 신경망을 인체 행동 인식, 제스처 인식 분야에 적용한 연구가 발표되었으며 비디오 영상과 같은 시계열 데이터 분석에 있어 2D 컨볼루션 신경망보다 더 적합하다는 것이 확인되었다[20, 21, 23, 24]. 본 연구에서는 열 화상 카메라로 촬영된 비디오 영상에 적합한 3D 컨볼루션 신경망 구조를 설계하고 낙상 인식 성능 향상을 이루고자 하였다.

3.1 3D 컨볼루션 및 풀링 연산

일반적으로 2D 컨볼루션 신경망은 단일 이미지를 입력으로 받아 컨볼루션 연산을 수행한다. 그에 반해 3D 컨볼루션 신경망은 일련의 이미지로 구성된 영상 볼륨을 입력으로 한다. 이로부터 각 이미지에 내포된 공간적 정보와 이전 이미지들과 현재 이미지 사이의 연속적인 시간적 정보를 함께 학습할 수 있다. 이러한 시공간적 특징 학습은 3D 컨볼루션 및 3D 풀링 연산을 통해 이루어진다. 본 논문에서 해결하고자 하는 낙상 인식 과제는 일정한 지속 시간 동안 순차적으로 발생하는 이벤트를 각 행동별로 분류하는 것으로 3D 컨볼루션 신경망을 통해 효과적으로 시공간적 정보를 추출할 수 있다.

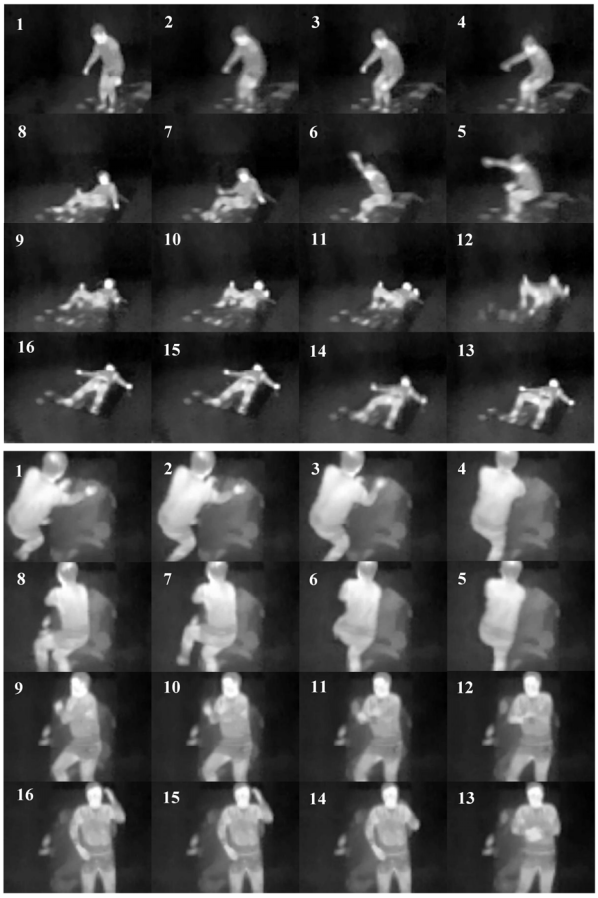
3D 컨볼루션은 다수의 인접한 프레임을 하나로 연결해 생성된 입력 볼륨에 대해 3D 커널을 적용한 연산을 수행한다. ReLU를 활성화 함수로 사용하는 i번째 층에 속한 j번째 특징 맵에서 (x, y, z) 위치의 값은 식 (1)에 의해 계산된다.

$$v_{ij}^{xyz} = \max\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}, 0\right) \quad (1)$$

이 때, b_{ij}은 바이어스 성분이며, P_i, Q_i는 각각 3D 커널의 높이, 너비 크기이며, R_i은 3D 커널의 시간 도메인에 대한 크기이다. w_{ijm}^{pqr}은 이전 층의 m번째 특징 맵과 연결된 (p, q, r) 위치의 가중치 값이다.

3.2 신경망 구조 및 학습

입력 볼륨 생성: 일반적으로 비디오의 길이에 따라 데이터의 양은 아주 방대해 진다. 하지만 짧은 시간 변화에 대해 영상이 근소한 차이를 보인다면 모든 연속적인 프레임들을 사용하는 것은 연산량 측면에서 매우 비효율적이며 실시간으로 낙상 인식을 위한 목표에 적합하지 않다. Schindler[25]의 연구에 따르면, 전체 40-120의 프레임 길이를 가지는 영상에 대해 7 프레임 정도의 아주 짧은 프레임 만으로도 기본 행동 인식을 수행하는데 충분하다는 것을 확인하였다. 본 논문에서는 입력 볼륨 생성을 위해 입력 영상에서 움직임이 두드러지게 발생한 주요 프레임 을 추출하였다. 키 프레임 추출은 인접한 두 프레임 간의 Optical flow를 계산하여 크기 값이 일정 임계값보다 큰 경우 움직임이 발생한 것으로 고려하여 키 프레임으로 선정하는 과정을 거쳤다. 전체 키 프레임의 길이는 실험을 통해 16 으로 설정하였으며 입력 볼륨의 차원은 3×16×112×112 (channels×length×height×width) 이다.



[Fig. 1] Example results of key-frame extraction. 'Falls from standing' action and 'Lying down' action

[Table 1] Network architecture of proposed model

Layer	Type	Filter size ($l \times h \times w$)	Output ($l \times h \times w$ @ v)
-	Input	-	16×112×112
1	Convolution Pooling	1×3×3 1×2×2	16×110×110@64 16×55×55@64
2	Convolution Pooling	1×3×3 1×2×2	16×53×53@128 16×27×27@128
3	Convolution Convolution Pooling	3×3×3 3×3×3 2×2×2	16×27×27@256 16×27×27@256 8×14×14@256
4	Convolution Convolution Pooling	3×3×3 3×3×3 2×2×2	8×14×14@512 8×14×14@512 4×7×7@512
5	Convolution Pooling	2×2×2 3×3×3	3×6×6@512 1×2×2@512
6	Fully connected	1×2×2	1×1×1@1024
7	Fully connected	1×1×1	1×1×1@1024
8	Softmax	-	1×1×1@14

네트워크 구조: 제안하는 3D 컨볼루션 신경망은 7개의 컨볼루션 층, 5개의 Max 풀링 층 및 2개의 완전 연결층과 Softmax 출력층으로 구성된다. 1, 2 layer 에서는 시간적인 정보를 유지하기 위해 커널 크기 1×3×3, stride 1×1×1의 컨볼루션 및 커널 크기 1×2×2, stride 1×2×2의 풀링으로 layer를 구성하였다. 3, 4 layer는 커널 크기 3×3×3, stride 1×1×1의 컨볼루션 및 커널 크기 2×2×2, stride 2×2×2의 풀링을 적용하였다. 5 layer에서는 커널 크기 2×2×2, stride 1×1×1 의 컨볼루션 및 커널 크기 3×3×3, stride 3×3×3의 풀링으로 설계하였다. 모든 풀링 연산은 Max pooling을 적용하였다. 두 개의 완전 연결층은 각각 1024개의 출력 유닛을 가진다. 네트워크 구조에 대한 세부 내용은 [Table 1]과 같다.

신경망 학습: 네트워크 학습은 낙상 인식을 위해 자체 수집한 데이터베이스에 대해 이루어졌다. 입력 데이터에 대해 각 모서리 및 센터 지점(×5)에 대해 잘라내기 및 프레임 간격 크기 2로 현재 키 프레임 및 전후의 프레임(-2, 0, +2)을 추가 선택(×3)하여 데이터 증대 작업을 수행하였다. 학습 파라미터는 Batch size 10, Learning rate 0.0003, Step size 5,000, Iteration 10,000 으로 설정하였으며, 확률 기울기 하강(Stochastic gradient descent) 알고리즘을 이용하여 학습하였다.

4. 실험 및 결과

본 장에서는 낙상 인식 실험을 위해 자체 구축한 데이터베이스에 대한 설명과 기존 모델과 제안한 모델을 이용한 인식 성능에 대해 비교 및 분석한다. 실험은 Intel(R) Xeon(R) E5-2650 v3 @ 2.30GHz CPU 및 NVIDIA Tesla K40c GPU 사양의 환경에서 수행되었다.

4.1 카메라 팬틸트 제어

열화상 카메라가 보다 넓은 영역을 모니터링 할 수 있도록 관심 영역을 실시간으로 추종하는 팬틸트 제어를 수행하였다. 이미지 내 환자 영역의 검출과 추적의 두 단계가 반복적으로 수행되며 카메라에 설치된 2개의 액추에이터를 제어하도록 하였다.

환자 영역의 검출은 열 영상 내 환자 영역과 주변 배경과의 온도 차를 이용한 임계값 기준으로 검출되며 이를 바탕으로 실시간 환자 추적을 수행한다. 환자 추적을 위한 알고리즘은 [Table 2]와 같으며, 기본적으로 이미지 내에서 검출된 환자 영역의 중심 $\boldsymbol{R}$과 카메라 이미지의 중심 $\boldsymbol{I}^{center}$ 사이의 거리 $\boldsymbol{D}$를 계산하며 미리 정의된 영역 $\boldsymbol{S}$를 유지하도록 동작한다. 만약 측정된 거리가 안전 영역 $\boldsymbol{S}$를 벗어난다면 시스템은 카메라에 연결된 액추에이터 제어를 통해 거리 $\boldsymbol{D}$를 조정한다.

4.2 *TCL Fall Detection Dataset*

일반적으로 대량의 데이터를 이용한 지도 학습을 통해 심층 신경망을 학습시킨다. 하지만 [Table 3]에 나타난 바와 같이 현재 낙상 인식을 위한 열 영상으로 구성된 양질의 공용 데이터베이스가 부족한 상황이다. 이를 보완하기 위해 24-36세의 성인 남녀 13명을 대상으로 낙상과 일반 행동으로 구성된 *TCL Fall Detection Dataset**을 구축하였다.

실험에 사용된 장비는 FLIR사의 AX8 열화상 카메라로 80×60 해상도를 가지는 소형화 모델로써 실시간 동영상 스트리밍이 가능하여 연속적인 상태 모니터링에 적합하다.

실험 데이터는 보행 중 낙상과 침실에서의 낙상의 두 가지 상황에 대해 촬영되었으며 낙상 421개, 일반 행동 831개로 구성된 총 1,252개의 비디오 영상으로 이루어졌다. 실험을 위한 병실 환경 및 카메라 설치는 [Fig. 2]에 나타난 바와 같이 구성하였다. 보행 중 낙상의 경우, 카메라와 피험자 사이의 거리는 3-4 m이며 카메라는 바닥에서 2 m 높이에 수평면 기준으로 13.5° 아래 방향으로 설치하였다. 침실에서의 낙상의 경우, 카메라와 피험자는 2.2-2.3 m 거리를 두고 있으며 카메라는 2.75 m 높이의 천장에 수직 아래 방향으로 부착하였다. 병실 침대의 높이는 0.5 m 이다. 데이터베이스는 2가지의 낙상, 12가지의 일상 생활 상황으로 구성된다. 각 상황은 실제 병실에서 자주 발생하는 전방, 후방, 측방 낙상과 병실 내 일상 생활에서 일어나는 행동으로 정의하였으며[5] 세부 내용은 [Table 4, Fig. 5, Fig. 6]와 같다. 각각의 행동들은 카메라 시점의 변화를 주며 촬영되었으며 이를 통해 수평면 360° 전방위에 대한 다각적인 행동 패턴을 데이터베이스에 포함시키고자 하였다.

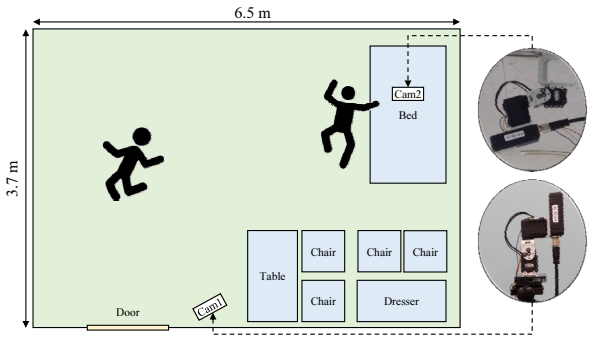
[Table 2] Human tracking algorithm with pan-tilt control

1:	Definition:
2:	Center of current detected region $\boldsymbol{R} = (R_x, R_y)$, Center of image $\boldsymbol{I^{center}} = (I_x^{center}, I_y^{center})$, distance $\boldsymbol{D} = \sqrt{(R_x - I_x^{center})^2 + (R_y - I_y^{center})^2}$, and the size of stable zone $\boldsymbol{S} = 0.25 * \boldsymbol{I}$
3:	**while** human is detected in the video sequence **do**
4:	Update current detected region center $\boldsymbol{R}$
5:	Calculate distance $\boldsymbol{D}$
6:	**if** $\boldsymbol{D} > \boldsymbol{S}$ **then**
7:	Transform the value of pan and tilt from 2D pixel coordinate to 3D world coordinate
8:	**end if**
9:	**end while**
10:	**Output**: The value of pan and tilt control p^w, t^w

[Table 3] Summary of existing thermal fall detection dataset

Reference	Sixsmith[26]	Kido[27]	Wong[14]	Vadivelu [16]
Year	2005	2009	2010	2016
Number of classes	2	2	4	-
Number of clips	44	400	50,000 images	44
Number of actors	1	5	1	-
Resolution	16×16	47×48	320×240	80×60
Available	×	×	×	O

* Available at https://drive.google.com/drive/folders/1HbizSUBOdZMTbzSjnppBK26eUSdxRj3v?usp=sharing



[Fig. 2] Experimental setup for data collection

[Table 4] 14 classes of *TCL Fall Detection Dataset*

Type	Classes
Falls	Falls from standing
	Falls from bed
Activities of daily living	Sitting down
	Standing up
	Object picking
	Walking
	Jogging
	Standing still
	Sitting on a bed
	Getting off a bed
	Moving on a chair
	Drinking
	Lying down
	Lying still

4.3 실험 결과 및 분석

제안한 모델의 성능 검증을 위해 *TCL Fall Detection Dataset*에 대해 기존 모델과 성능 비교 실험을 진행하였다. 일반적인 Hand-crafted 특징 기반 모델과 컨볼루션 신경망 구조의 성능 비교를 통해 낙상 인식 연구에 있어 심층 신경망의 성능 유효성을 검증하고, 단일 프레임 처리 방식의 2D 신경망 모델과 제안한 3D 컨볼루션 신경망 모델과의 성능 비교를 통해 시계열 데이터에 대해 3D 컨볼루션 연산을 통한 Spatiotemporal 특징 학습이 성능 향상에 두드러지는 영향을 보이는지 검증하고자 하였다. 입력 데이터는 3.2장에서 설명한 바와 같이 전처리 단계에서 추출된 16개의 키 프레임을 입력 볼륨(3×16×112×112)으로 하였다.

실험 성능의 Base line 으로 단일 프레임 단위로 환자 영역에 대해 HOG 특징을 추출하고 linear SVM을 사용하여 분류 작업을 수행하였다. 전처리 작업으로 추출된 키 프레임에 대해 온도 임계치 기반 배경 제거 방법을 적용하여 환자 영역을 검출한 후, 검출된 환자 영역에 대해 HOG 특징 벡터를 추출하였다. HOG의 셀 크기는 8×8 에 대해 히스토그램을 계산하였다. 또한 현재 비디오 데이터에서 추출된 Hand-crafted 특징 중에서 최고의 인식 성능을 보이는 것으로 알려진 iDT[19] 을 이용한 모델에 대해서도 실험을 수행하였다. 16개의 입력 프레임에서 검출된 환자 영역에 대해 HOG, HOF, MBHx, MBHy 국부 특징을 계산하고 각 채널에 대해 Fisher vector로 인코딩한 특징 벡터를 L-2 정규화 과정을 거친 후 하나로 연결하여 특징 벡터화 하였다. 인식 성능을 측정하기 위해 다중 클래스 linear SVM을 사용하였다. 컨볼루션 신경망 구조에 대한 인식 성능을 검증하기 위해 수백만 장의 이미지 데이터베이스인 ImageNet을 이용하여 신경망을 학습시킨 2D 컨볼루션 신경망 구조[28]에 대해 낙상 인식 데이터베이스에 대한 실험을 수행하였다. 입력 데이터는 전처리 단계에서 추출된 16개의 키 프레임을 입력 데이터로 넘겨 주었다. 단일 프레임 단위 처리 방식으로 각 프레임에 대해 독립적인 클래스별 확률을 계산하고 그 값을 합산하여 가장 높은 값을 얻은 클래스를 해당 Clip의 최종 인식 결과로 산출하는 방식을 사용하였다. C3D[21]은 비디오 분석을 위해 고안된 3D 컨볼루션 신경망 모델이며 8개의 3D 컨볼루션 층, 5개의 3D 풀링 층과 2개의 완전 연결층으로 이루어진 3D 컨볼루션 신경망 구조를 가지고 있다. 3D 컨볼루션의 커널 크기는 3×3×3이며, 1×2×2 및 2×2×2 커널 크기의 Max pooling 으로 구성된다. 두 개의 완전 연결층은 4096개의 Output unit을 가진다.

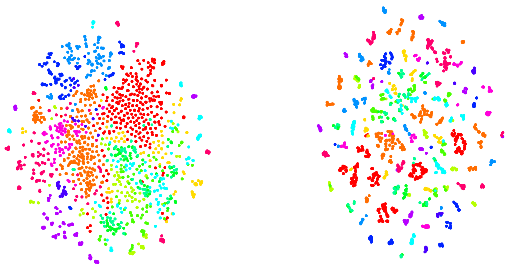
인식 성능 평가를 위해 모든 비교 모델에 대해 Leave-One-Subject-Out 방식을 사용하였다. 전체 13명의 피험자 중, 1명의 피험자를 제외한 나머지 피험자의 데이터를 이용하여 모델을 학습시키고 남겨진 1명의 피험자의 데이터에 대해 테스트를 진행하는 방식을 각각의 피험자에 대해 수행하여 평균 인식 성능을 계산하였다. 각 모델에 대한 비디오 Clip 평균 인식 성능은 [Table 5]와 같다.

실험 결과 본 논문에서 제안한 3D 컨볼루션 신경망 모델이 다른 4가지 모델과 비교하여 우수한 성능을 보임을 확인하였다. 제안한 모델은 실험의 Base line인 HOG 및 SVM 을 이용한 모델과 비교하여 8.9%, iDT 정보와 Fisher vector 를 통해 특징 벡터화한 방식의 모델에 비해서는 7.2% 높은 성능을 나타내었다. 단일 프레임 처리 기반의 ImageNet 신경망 모델은 63.2%의 인식 성능을 보였는데, 이는 클래스 간의 데이터 Variance가 크지 않은 실험 데이터베이스에 대해 인접 프레임 사이의 시간적 관계를 고려하지 않으면 변별력 있는 행동 분류를 하기 힘듦을 나타낸다. 또한 C3D 모델보다 약 7.6배 더 가벼운 네트워크 구조를 사용했음에도 90.2%로 1.3% 높은 인식 성능을 보였다.

특징 임베딩: 제안한 3D 컨볼루션 신경망을 이용하여 학습된 특징들이 다른 클래스와 구분되는 변별력 있는 정보를 포함하고 있는지를 확인하기 위해 [Fig. 3]와 같이 HOG + linear SVM 모델과 제안한 모델의 특징 임베딩 가시화 결과를 비교하였다. 각각의 특징 가시화는 데이터 증대 처리를 한 데이터베이스에 대해 임의의 9,000개의 영상 클립을 선정하고 2차원 평면에 투영시키는 과정을 거쳤다. 제안한 모델에 대해 fc6

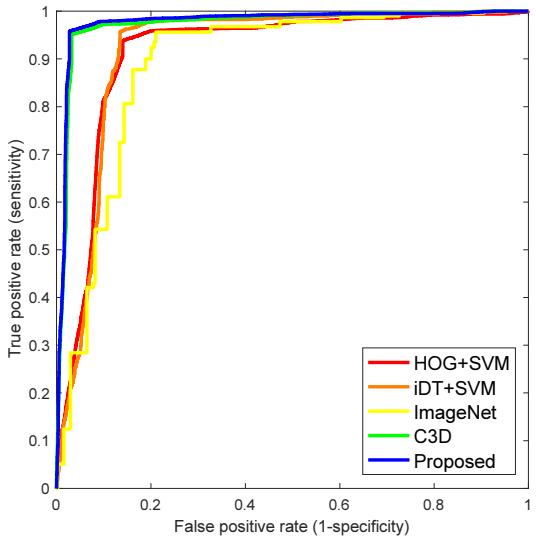
[Table 5] Experimental result on *TCL Fall Detection Dataset*

Method	Clip Accuracy (%)
HOG + linear SVM	81.3
iDT w/ Fisher vector + linear SVM[19]	83.0
Single-frame CNN[28]	63.2
C3D[21]	88.9
Proposed	**90.2**



(a) HOG + linear SVM model (b) Proposed model

[Fig. 3] Feature embedding visualizations of HOG + linear SVM model and proposed model on *TCL Fall Detection Dataset* using t-SNE[29]



[Fig. 4] Comparing ROC curves

layer에서 추출된 특징들을 사용하였으며 각각의 영상들은 1개의 데이터 포인트로 표현되며 같은 클래스에 속하는 데이터 포인트는 같은 색깔을 가진다. 3D 컨볼루션 신경망에 의해 추출된 특징이 HOG 특징에 비해 좀 더 밀집된 클래스별 군집을 형성하며 분포하는 것을 확인할 수 있다.

ROC 곡선 분석: 실제 병실 환경에서는 낙상 이벤트에 대해 낙상 인식 시스템이 낙상으로 감지하는 정도를 나타내는 민감도(Sensitivity)와 일상 생활 속 행동들에 대해 정상으로 감지하는 정도인 특이도(Specificity)가 낙상 인식 시스템의 중요한 성능 지표로 작용한다. 제안하는 낙상 인식 모델의 민감도와 특이도 간의 관계를 분석하기 위해 ROC(Receiver operating characteristic) 곡선을 [Fig. 4]에 그래프로 나타내었다. 이 때, 앞서 정의된 14개의 클래스에 대해 낙상과 일반 행동의 두 가지 그룹으로 나누어 ROC 곡선을 표현하였다. 각 모델의 AUC(Area under the curve) 값은 0.912, 0.919, 0.890, 0.973, 0.977 로 측정되었다. 제안하는 모델의 경우 1에 가장 가까운 수치를 보여 다양한 상황에서 더욱 정확한 낙상 인식이 가능함을 확인하였다. 또한 [Fig.7]에 나타낸 바와 같이 제안된 모델에 대한 Confusion matrix 분석에서 'Walking'과 'Jogging' 클래스에 대한 오인식이 빈번하게 일어나는 한계가 있음을 확인하였다. 이는 본 연구에서 적용한 키 프레임 추출 기법이 인접 프레임 간의 움직임에 기반하여 일정한 길이로 생성되므로 실제 시간 간격에 대한 차이를 고려하지 못하기 때문이나 환자 안전에 직결되는 낙상 인식에는 큰 영향을 미치지 않는다.

5. 결 론

본 연구에서는 병실 내 환자의 낙상사고 인식을 위한 시계열 데이터 특징 학습 문제를 3D 심층 신경망 학습 기법을 적용하여 해결하고자 하였다. 제안한 3D 컨볼루션 신경망 구조는 키 프레임 추출 단계에서 생성된 입력 볼륨을 7개의 3D 컨볼션층, 5개의 3D Max 풀링 층, 2개의 완전 연결층 및 Softmax 출력층으로 구성된 네트워크 구조를 거쳐 시계열 영상 데이터에 내포된 낙상 이벤트의 시공간적 정보를 효과적으로 학습 가능하도록 하였다. 성능 평가를 위해 자체 구축한 *TCL Fall Detection Dataset* (available at https://drive.google.com/drive/folders/1HbizSUBOdZMTbzSjnppBK26eUSdxRj3v?usp=sharing)에 대해 다른 모델과 비교 실험함으로써 제안한 모델이 높은 성능을 보임을 검증하였다.

제안한 3D 컨볼루션 신경망을 적용한 낙상 모니터링 시스템의 주야간 감시를 통해 낙상 사고에 대한 빠른 대응을 가능케할 수 있다. 또한 의료 기관에서는 환자수 대비 부족한 간호 인력에 대한 해결 방안으로써 보다 높은 질의 간호 서비스를 제공할 수 있을 것이다. 또한, 현재 구축된 데이터의 한계로 낙상이 아닌 발작 등 다른 응급 상황의 모니터링이 어려우나 추후에 여러 응급상황에 대한 데이터베이스를 구축할 경우 병실 내 환자의 안전을 보호하는 중요한 역할을 기대할 수 있을 것이다.

References

[1] A. L. Hendrich, P. S. Bender, and A. Nyhuis, "Validation of the Hendrich II Fall Risk Model: A large concurrent case/control study of hospitalized patients," *Applied Nursing Research*, vol. 16, no. 1, pp. 9-21, February, 2003.

[2] A. Kalache, D. Fu, S. Yoshida, W. Al-Faisal, L. Beattie, W. Chodzko-Zajko, H. Fu, K. James, S. Kalula, B. Krishnaswamy, N. Kronfol, P. Marin, I. Pike, D. Rose, V. Scott, J. Stevens, C. Todd, and G. Usha, "World Health Organisation Global Report on Falls Prevention in Older Age," Geneva: WHO 2007, World Health Organization, Department of Ageing & Life Course, ed., 2007.

[3] L. Z. Rubenstein, "Falls in older people: epidemiology, risk factors and strategies for prevention," *Age and Ageing*, vol. 35, no. suppl_2, pp. ii37-ii41, 2006.

[4] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, no. 16, pp. 144-152, 2013.

[5] J. G. Shin, "Prevalence and management of falls among elderly in a hospital," *Unpublished master's thesis,* Hanyang University, pp. 1-63, 2011.

[6] H. Rimminen, J. Lindström, M. Linnavuo, and R. Sepponen, "Detection of falls among the elderly by a floor sensor using the

electric near field," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 6, pp. 1475-1476, 2010.

[7] S. Shan and T. Yuan, "A wearable pre-impact fall detector using feature selection and support vector machine," *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING*, Beijing, China, 2010, doi: 10.1109/ICOSP.2010.5656840.

[8] B. Kaluza and M. Luštrek, "Fall detection and activity recognition methods for the confidence project: a survey," *10th International Multiconference*, Ljubljana, Slovenia, pp. 22-25, 2008.

[9] R. Igual, C. Medrano, and I. Plaza, "Challenges, issues and trends in fall detection systems," *BioMedical Engineering*, [Online], vol. 12, no. 1, pp. 66, DOI. Available: https://doi.org/10.1186/1475-925X-12-66.

[10] M. A. Habib, M. S. Mohktar, S. B. Kamaruzzaman, K. S. Lim, T. M. Pin, and F. Ibrahim, "Smartphone-Based Solutions for Fall Detection and Prevention: Challenges and Open Issues," *Sensors 2014*, vol. 14, no. 4, pp. 7181-7208, 2014.

[11] M. K. Mulcahy and S. Kurkovsky. "Automatic Fall Detection using Mobile Devices," *2015 12th International Conference on Information Technology*, Las Vegas, NV, USA, 2015, doi: 10.1109/ITNG.2015.98.

[12] Z. Fu, E. Culurciello, P. Lichtsteiner, and Tobi Delbruck, "Fall detection using an address-event temporal contrast vision sensor," 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, USA, 2008, doi: 10.1109/ISCAS.2008.4541445.

[13] S.-G. Miaou , P.-H. Sung, and C.-Y. Huang, "A customized human fall detection system using omni-camera images and personal information," *1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare*, Arlington, VA, USA, 2006, doi: 10.1109/DDHH.2006.1624792.

[14] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," *2010 Second International Conference on Computer Research and Development*, Kuala Lumpur, Malaysia, 2010, doi: 10.1109/ICCRD.2010.163.

[15] G. Mastorakis and D. Makris, "Fall detection system using Kinect's infrared sensor," *Journal of Real-Time Image Processing*, vol. 9, no. 4, pp. 635-646, December, 2014.

[16] S. Vadivelu, S. Ganesan, O. V. R. Murthy, and A. Dhall, "Thermal Imaging Based Elderly Fall Detection," *Computer Vision – ACCV 2016 Workshops*, Taipei, Taiwan, pp. 541-553, 2016.

[17] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *15th ACM international conference on Multimedia*, Augsburg, Germany, pp. 357-360, 2007.

[18] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," *British Machine Conference*, pp. 99.1-99.10, 2008.

[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," *2013 IEEE International Conference on Computer Vision*, Sydney, Australia, 2013, doi: 10.1109/ICCV.2013.441.

[20] S. Ji, W. Xu, M. Yang, and Kai Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan., 2013.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, doi: 10.1109/ICCV.2015.510.

[22] K. Simonyan and A. Zisserman. "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems 27*, Montreal, Canada, pp. 568-576, 2014.

[23] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Boston, MA, USA, 2015, doi: 10.1109/CVPRW.2015.7301342.

[24] M. Baccouche,F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," *International Workshop on Human Behavior Understanding*, 2011, doi: 10.1007/978-3-642-25446-8_4.

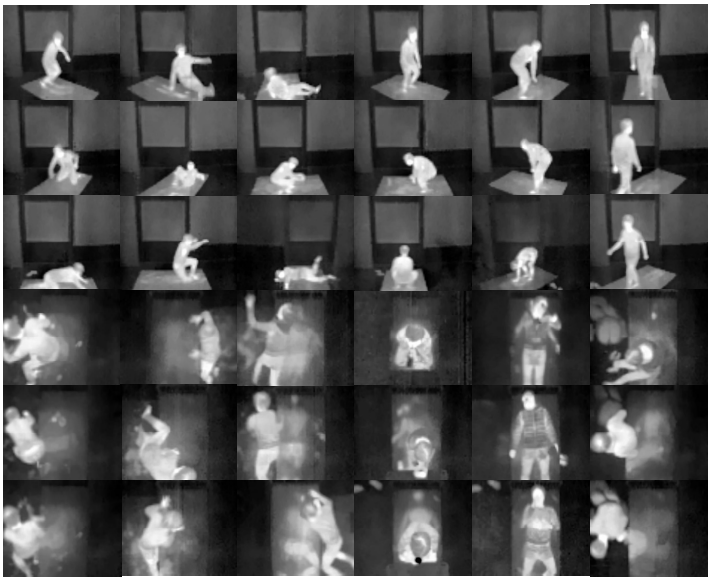
[25] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, doi: 10.1109/CVPR.2008.4587730.

[26] A. Sixsmith, N. Johnson, and R. Whatmore, "Pyroelectric IR sensor arrays for fall detection in the older population," *Journal de Physique IV*, Vol. 128, pp. 153-160, September, 2005.

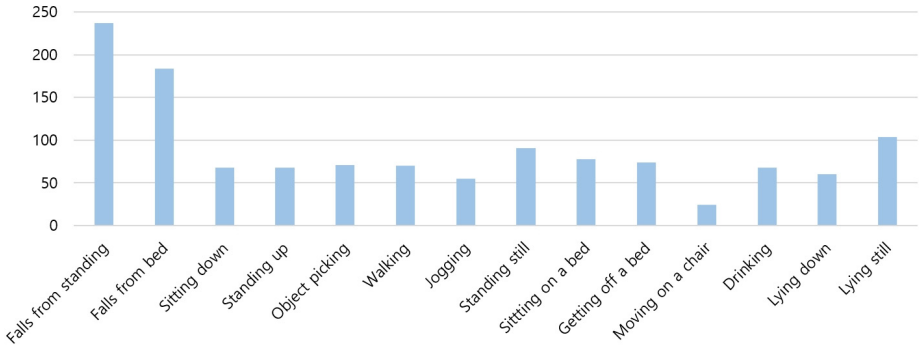
[27] S. Kido, T. Miyasaka, T. Tanaka, T. Shimizu, and T. Saga, "Fall detection in toilet rooms using thermal imaging sensors," *2009 IEEE/SICE International Symposium on System Integration*, Tokyo, Japan, 2009, doi: 10.1109/SI.2009.5384550.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, June, 2012.

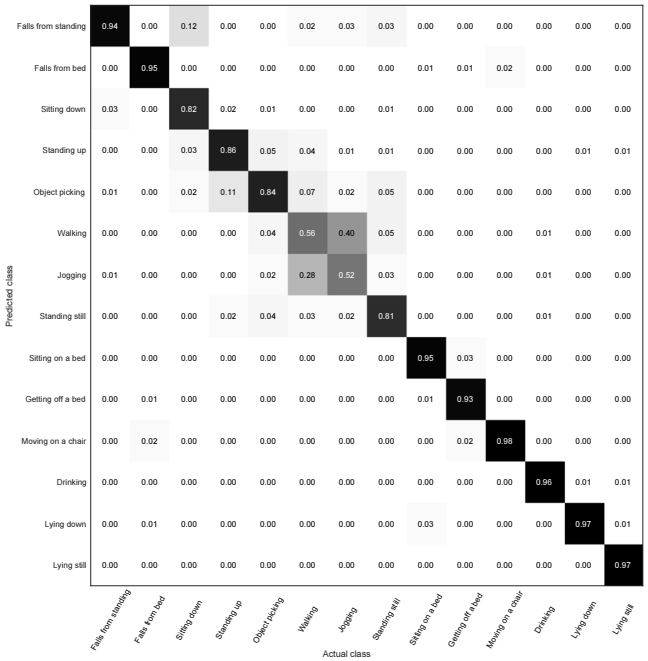
[29] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research 9*, pp. 2579-2605, Nov., 2008.



[Fig. 5] Example of the *TCL Fall Detection Dataset*. In all cases, situations are considered that occur while living in a real hospital. We also consider changes of view point to handle various human activity patterns. The upper 3 rows are selected from standing or walking situation and the lower 3 rows from lying in a hospital bed situation



[Fig. 6] Number of clips per action class



[Fig. 7] Confusion matrix by the proposed model

김 대 언

2014 경북대학교 전자공학부 (공학사)
2016 한국과학기술원 미래자동차학제전공 (공학석사)
2016 ~ 2017 인간로봇상호작용 핵심연구센터 연구원
2017 ~ 현재 유진기술(주) 연구원

관심분야: 영상처리, 기계학습

전 봉 규

2015 금오공과대학교 기전공학과 (공학사)
2017 한국과학기술원 로봇공학학제전공 (공학석사)
2017 ~ 현재 한국과학기술원 로봇공학학제전공 박사과정

관심분야: 로보틱스

권 동 수

1980 서울대학교 기계공학과(공학사)
1982 한국과학기술원 기계공학과 (공학석사)
1991 Georgia Institute of Technology 기계공학과 (공학박사)
1995 ~ 현재 한국과학기술원 기계공학과 교수

관심분야: Human-Robot Interaction, Haptics, Medical Robots