KIPS Tr. Software and Data Eng.
Vol.6, No.3 pp.155~160 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.3.155

A Method of Activity Recognition in Small-Scale Activity Classification Problems via Optimization of Deep Neural Networks

Seunghyun Kim† · Yeon-Ho Kim†† · Do-Yeon Kim†††

ABSTRACT

Recently, Deep learning has been used successfully to solve many recognition problems. It has many advantages over existing machine learning methods that extract feature points through hand-crafting. Deep neural networks for human activity recognition split video data into frame images, and then classify activities by analysing the connectivity of frame images according to the time. But it is difficult to apply to actual problems which has small-scale activity classes. Because this situations has a problem of overfitting and insufficient training data. In this paper, we defined 5 type of small-scale human activities, and classified them. We construct video database using 700 video clips, and obtained a classifying accuracy of 74.00%.

Keywords : Activity Recognition, Deep Neural Network, LRCN, Optimization

심층 신경망의 최적화를 통한 소규모 행동 분류 문제의 행동 인식 방법

김 승 현† · 김 연 호†† · 김 도 연†††

요 약

최근 컴퓨터를 이용한 다양한 인식 문제를 해결하기 위해 딥 러닝을 적용하는 사례가 늘어나고 있다. 딥 러닝은 학습에 필요한 요소를 학습 데이터를 통해 스스로 도출해내기 때문에, 수작업(hand-craft)을 통해 특징을 도출하던 기존의 기계학습 방법보다 더 많은 장점을 갖는다. 행동 인식을 위한 기존의 심층 신경망은 비디오 데이터를 일정 프레임의 이미지로 분할한 후, 분할된 각 이미지 사이의 시간적 연계성 분석을 통해 행동을 분류한다. 그러나 이러한 신경망은 소규모 행동 클래스를 갖는 분류 문제에서 학습 데이터의 부족 문제 및 과적합(overfitting) 문제로 인해 이를 실제 문제에 적용하기 어려운 경우가 많다. 이에 본 논문에서는 5가지의 소규모 행동 클래스를 정의하고, 기존 행동 인식 신경망의 최적화를 통해 이를 분류하였다. 700개의 비디오데이터를 통해 행동 데이터베이스를 구성하였고, 약 74.00%의 분류 정확도를 얻을 수 있었다.

키워드 : 행동 인식, 심층 신경망, LRCN, 최적화

1. 서 론

행동 인식은 관찰 대상의 동작이나 움직임에 대한 의미를 분별하고 판단하는 것을 말하며, 감시(surveillance), 로봇(robot), 인터페이스(interface) 등 다양한 응용 분야를 가진 핵심적인 기술 중 하나이다. 일반적인 행동 인식 과정은 미리 정의된 행동 모델을 기반으로 새로 입력된 행동 패턴을 분석하고 분류하는 것에 중심을 둔다[1]. 행동 패턴은 특정 행동에서 나타나는 움직임에 대한 추상적인 규칙이며, 행동 인식에 중요한 영향을 미치는 요소들의 집합이라고 할 수 있다. 따라서 영상을 통해 행동을 인식하려면 중요도가 높은 특징점들을 선정하여 해당 픽셀들의 특징 벡터를 표현할 수 있어야 하고, 일관성을 유지한 채 패턴을 비교할 수 있어야 한다. 하지만 수작업(hand-craft)을 통한 특징점 추출에는 한계가 있으며, 영상의 해상도가 높아질수록 고려해야할 특징 픽셀의 개수가 기하급수적으로 증가하므로 구현에 어려움이 따른다.

한편, 데이터를 학습하고 분류하기 위한 방법으로써 딥 러닝이 재조명되고 있다. 딥 러닝은 다수의 계산 층으로 구성된 신경망 모델을 통해 데이터의 패턴을 학습하며[2], 최근 발전된 컴퓨팅 성능과 알고리즘을 바탕으로 기존의 기계학습 영역에서 뛰어난 성능을 보이고 있다[3]. 또한 학습 데

※ This work was supported by the Nuclear Safety Research Program through the Korea Foundation Of Nuclear Safety(KOFONS), granted financial resource from the Nuclear Safety and Security Commission (NSSC), Republic of Korea (No.1403025).
† 준 회 원 : 순천대학교 컴퓨터비전 및 보안실험실 연구원
†† 준 회 원 : 순천대학교 컴퓨터과학과 석사과정
††† 비 회 원 : 순천대학교 컴퓨터공학과 교수
Manuscript Received : September 8, 2016
Accepted : October 27, 2016
* Corresponding Author : Do-Yeon Kim(dykim@sunchon.ac.kr)

이터를 통해 스스로 특징을 찾고 패턴을 학습하기 때문에 구현의 어려움을 완화시킬 수 있는 좋은 수단이다.

D. Jeffrey et al.[4]의 연구에서는 행동 인식을 위해 CNN (convolutional neural networks)과 LSTM (long short-term memory networks)을 이용한 모델인 LRCN (long-term recurrent convolutional networks)을 제안하였다. 이 신경망 모델은 비디오 클립의 프레임 이미지를 추출한 후, CNN을 통해 각 이미지의 특징을 획득하고, 획득된 특징을 다시 LSTM을 통해 학습시킴으로써, 시계열적 행동 패턴의 특징을 학습할 수 있도록 하였다. 그러나 기존 연구의 신경망은 101가지의 행동에 대한 분류를 수행하기 위해 가중치가 최적화되어 있기 때문에 소규모 행동 분류 문제의 행동 인식에 응용하기 어렵다. 가령, 침입자 행동 인식을 위한 분류 문제의 경우, 침입 행위로 판단될 수 있는 행동의 종류는 상기한 101가지보다 훨씬 적다. 학습 데이터의 클래스 개수가 감소하면, 과적합(overfitting)이 발생할 확률이 높아지기 때문에 결과적으로 분류 정확도가 현저히 떨어지게 된다.

따라서 본 논문에서는 기존 신경망의 최적화를 통해 클래스 개수가 적은 소규모 행동 분류 문제의 행동 인식을 위한 방법을 제안한다. 실험 범위 설정 및 결과 도출을 위해 원전에서 일어날 수 있는 5가지 침입행동 유형[5]; 달리기(running), 걷기(walking), 기어가기(crawling), 점프하기(jumping), 기어오르기(climbing)로 제한하였으며, 딥 러닝 구동을 위해 카페(caffe) 프레임워크를 사용하였다.

논문의 2장에서는 행동 인식과 관련된 연구현황과 기존의 LRCN 모델에 대하여 기술하였고, 3장에서는 인간 행동 인식을 위한 신경망에 대하여 기술하였다. 4장에서는 해당 신경망의 학습·분류 실험 및 성능평가에 대하여 기술하였고, 마지막 5장에서는 결론 및 향후과제에 대하여 기술하고 마무리하였다.

2. 관련 연구

2.1 행동 인식 연구현황

행동 인식은 지난 몇 년 동안 다양한 접근 방법과 방법론을 통해 연구되어 왔다. 연구 부류는 두 가지 형태로써 구분될 수 있는데 행동을 여러 가지 특성으로 나누고 카테고리를 만들어 행동 자체를 연구한 연구 부류와, 행동을 수행하는 대상에 대하여 분석과 확률적 측면을 바탕으로 행동을 인식하려는 연구부류가 있었다[6]. 본 논문은 행동 자체를 연구하는 것이 아닌, 어떤 행동에 대한 데이터가 주어졌을 때, 그것을 영상 분석과 확률적 측면에서 분류하는 것이 목표이므로 이러한 연구 부류에 대한 언급은 생략한다.

대부분의 선행 연구들은 카메라를 통해 관찰되는 영상 기반 데이터들을 이용하거나, 센서를 통해 측정된 데이터를 이용하였다. 분류에 사용된 알고리즘으로는 기계학습 관련 알고리즘이나 논리 모델 및 추론 알고리즘들이 많았다.

2011년에 진행된 한 행동 인식 연구 동향[7]에 의하면, 비전과 센서를 기반으로 한 기계학습 알고리즘이 행동 인식 방법의 주류를 이루었다. 기계학습 알고리즘은 지도 학습(supervised learning) 알고리즘과 비지도 학습(unsupervised learning) 알고리즘이 사용되었으며, 종류로는 HMM(hidden Markov model), dynamic and naive bayes networks, decision tree, SVM (suport vector machine) 등이 있다.

반면, 2015년에 진행된 다른 동향 연구[8]에서는 앞서 언급한 동향 연구와 방법론적 측면에서 크게 다른 흐름은 없었지만, 딥 러닝(deep learning)을 이용한 사례가 눈에 띄게 늘었다. 딥 러닝은 학습 데이터를 통해 유용한 특징을 직접 찾고 학습하며, 기존 알고리즘보다 향상된 성능을 보이기 때문이다[9].

2.2 Long-term Recurrent Convolutional Networks

LRCN 모델은 비디오 영상으로부터 사람의 행동을 학습하고 분류하기 위해 고안된 심층 신경망 모델이다. 이 모델은 행동 학습을 위해 일정 시간 간격의 프레임(RGB frame)들과 해당 프레임들 사이의 옵티컬 플로우(optical flow)를 이용하므로, 이를 추출하기 위한 전처리 과정을 거친다. 추출된 이미지들은 4가지 신경망 모델; (a) singleFrame_RGB, (b) singleFrame_flow, (c) lstm_RGB, (d) lstm_flow의 입력으로 사용된다. 출력 결과는 각 신경망의 고유한 분류 결과 외에도, (a)와 (b), (c)와 (d) 신경망의 혼합된 형태의 계산 결과까지 추가하여, 총 8개의 분류 결과가 출력된다. 출력의 세부 내용은 Table 1과 같다.

Table 1. Details Type of the Outputs in LRCN

No.	Network	Fusion rate of Outputs
1	(a) singleFrame_RGB	(a) = 1
2	(b) singleFrame_flow	(b) = 1
3	(c) lstm_RGB	(c) = 1
4	(d) lstm_flow	(d) = 1
5	×	(a) : (b) = 0.5 : 0.5
6	×	(a) : (b) = 0.33 : 0.67
7	×	(c) : (d) = 0.5 : 0.5
8	×	(c) : (d) = 0.33 : 0.67

LRCN의 가장 큰 특징은 비디오 영상 학습을 위해 CNN과 LSTM을 조합하여 사용한다는 것이다. 일반적으로 CNN은 정지 영상인 이미지의 학습 및 분류에 많이 사용되고, LSTM은 자연어 처리와 같은 시계열 데이터 학습에 많이 사용된다. 하나의 비디오 영상은 시간에 따라 변화하는 연속된 이미지들의 모임이라고 할 수 있으므로 CNN과 LSTM의 조합은 비디오 영상 학습에 적합하다.

3. 행동 인식을 위한 심층 신경망의 최적화

3.1 학습을 위한 행동 데이터베이스의 구성

행동 데이터베이스의 세부사항은 Table 2와 같으며, 모든 영상은 유튜브(youtube) 검색 결과를 바탕으로 학습 능률 향상을 위해 가로 640 pixel, 세로 480 pixel의 사이즈를 갖도록 녹화하였다. 행동 데이터베이스는 행동 유형별로 140개씩 총 700개의 영상으로 구성된다.

Table 2. Details of the Activity Database

Activity	Training	Validation	Test
Climbing	80	20	40
Crawling	80	20	40
Running	80	20	40
Walking	80	20	40
Jumping	80	20	40

각 행동 유형에 따른 영상 데이터는 학습(training), 검증(validation), 테스트(test)와 같은 3가지 유형으로 분류된다. 학습 데이터셋은 80개 중에서 30~35개 영상은 근거리, 25~30개 영상은 원거리 촬영 영상이며, 나머지 영상은 기존 영상을 플립(flip) 시키거나, 채도, 명암 수치를 수정한 영상들로 채워 넣었다. 다음 Fig. 1은 사용된 영상데이터 중 crawling 행동을 나타내고 있는 데이터의 예이다.

Fig. 1. Example of Activity Data (Crawling)

3.2 신경망 최적화 및 세부 파라미터 조정

기존 신경망 모델은 101가지의 행동 클래스를 분류하도록 구성되어 있기 때문에 최적화 과정을 거치지 않고 훈련시키게 될 경우, 훈련에 상당한 시간이 소요되며, 데이터 부족 및 적은 클래스 개수로 인해 원하는 정확도를 얻기 힘들다. 따라서 5가지 행동 분류에 적합한 계산을 위해 해당 신경망을 최적화하고, 세부 파라미터를 수정한다. 최적화는 상기한 네 가지 신경망 모델에 각각 적용된다.

먼저 singleFrame_RGB, singleFrame_flow 모델의 최적화를 위해 선행 학습된 네트워크 모델의 가중치를 반영하였다. singleFrame의 두 네트워크 모델은 유사한 형태로 구성되어 있으며, 네트워크 구조는 Fig. 2와 같다. 그림에 나타난 수치는 해당 블록의 반복 횟수를 나타낸다.

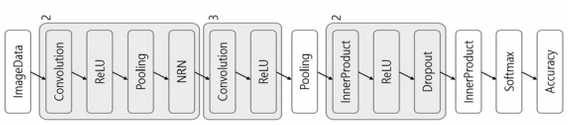


Fig. 2. Network Architecture of singleFrame RGB and Flow

singleFrame RGB and flow 네트워크는 전형적인 CNN 모델과 비슷한 구조이며, 비디오 프레임의 특징만을 학습하고, 시계열적 특성에 대한 학습은 진행되지 않는다. 해당 모델의 학습을 진행할 때, 목적에 부합하지 않는 잘못된 데이터로 인한 가중치의 오염을 방지하기 위해 특정 레이어의 lr_mult와 decay_mult 값을 프리징(freezing) 시켰으며, 마지막 fully-connected 레이어의 num_output 값을 101에서 5로 수정하였다.

lstm_RGB 모델과 lstm_flow 모델의 최적화도 위와 비슷한 순서로 진행된다. 두 모델에는 각각 50,000번, 5,000번의 학습을 반복한 선행 학습 모델을 사용하였다. lstm의 두 네트워크 모델은 유사한 형태로 구성되어 있으며, 레이어 구조는 Fig. 3과 같다.

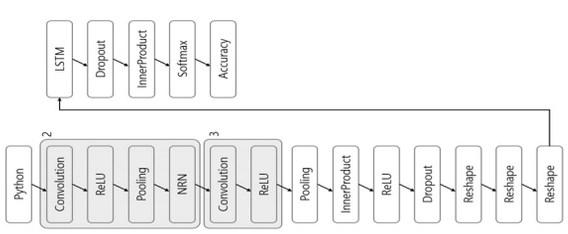


Fig. 3. Network Architecture of lstm RGB and Flow

lstm RGB and flow 모델은 singleFrame RGB and flow 모델보다 약간 더 복잡한 구조를 갖는다. lstm RGB and flow 모델은 CNN을 거쳐 특징맵을 추출한 후, LSTM을 통해 시계열적 특성까지 학습한다. singleFrame 모델과 마찬가지로 가중치의 오염을 방지하기 위해 특정 레이어의 lr_mult와 decay_mult 값을 프리징 시키고, 마지막 fully- connected 레이어의 num_output 값을 5로 낮춘 후 최적화를 진행하였다. Table 3은 Solver 파일들의 세부사항을 나타낸다. (a)~(d)는 singleFrame_RGB, singleFrame_flow, lstm_RGB, lstm_flow 네트워크를 의미한다.

Table 3. Details of Solver Files

Network	Max Iteration	Step Size	Momentum	Weight Decay	Base Learning Rate
(a)	3,000	3,000	0.9	0.005	0.001
(b)	16,000	20,000	0.9	0.005	0.001
(c)	10,000	10,000	0.9	0.005	0.001
(d)	10,000	20,000	0.9	0.005	0.001

4. 실험 및 성능평가

4.1 신경망 훈련 및 검증

행동 분류에 앞서 신경망이 제대로 최적화 되었는지 확인하기 위해 검증 데이터셋을 이용해 행동 분류를 수행하였다. 이 테스트의 목적은 신경망이 학습을 제대로 진행 하였는지 확인하는 것이다. 다음 Table 4는 검증 데이터셋을 이용하여 분류를 진행 하였을 때, 분류 결과를 나타낸 것이다.

Table 4. Classification Result of Validation Dataset

Activity	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Climbing	3	4	18	20	3	5	19	19
Crawling	3	4	17	12	4	3	18	18
Running	5	2	18	19	4	4	19	19
Walking	1	5	20	18	1	1	20	20
Jumping	7	4	15	20	7	7	18	20

위 Table에서 (a)~(d)는 각각 (a) singleFrame_RGB, (b) singleFrame_flow, (c) lstm_RGB, (d) lstm_flow를 의미하며, (e)~(h)는 (a)~(d) 네트워크들을 0.5 : 0.5 또는 0.33 : 0.67의 비율로 혼합한 네트워크를 의미한다(세부내용은 Table 1에서 기재). 각 네트워크에 입력된 검증 데이터 개수는 행동별로 20개씩 사용되었으며, Table 4에 나타난 숫자는 검증 데이터 20개 중에서 정확하게 분류해낸 데이터의 개수를 의미한다.

전체적인 데이터 분류 결과를 분석해 보면, 시계열 데이터를 학습할 수 있는 네트워크인 (c), (d), (g), (h)의 분류 결과가 (a), (b), (e), (f) 네트워크의 분류 결과보다 훨씬 분류 성능이 좋은 것을 확인할 수 있다. 다음 Table 5는 분류 결과에 대한 정확도를 나타낸 것이다. Table에서 CNN은 (a), (b), (e), (f) 네트워크의 종합적인 결과를 의미하며, LRCN은 (c), (d), (g), (h) 네트워크를 의미한다.

Table 5. Classification Accuracy of Validation Dataset

Activity	CNN	LRCN
Climbing	18.75%	95.00%
Crawling	17.50%	81.25%
Running	18.75%	93.75%
Walking	10.00%	97.50%
Jumping	31.25%	91.25%
Total	19.25%	91.75%

만약 신경망이 과적합 상태에 도달하였다면, 네트워크 종류에 상관없이 모든 결과가 높은 분류성능을 보여주어야 한다. 따라서 위 실험에서 해당 신경망이 과적합 상태에 도달했다고 보긴 어려우며, 행동 패턴에 대한 일반성을 갖고 있다고 볼 수 있다.

4.2 신경망 성능검증 및 평가

검증 데이터셋은 신경망이 행동 데이터베이스를 학습할 때, 가중치 조절에 영향을 미치는 데이터셋이기 때문에 해당 분류 정확도를 모두 신뢰하긴 어렵다. 따라서 별도의 테스트 데이터셋을 통해 신경망을 검증할 필요가 있다. 다음 Table 6은 구성된 테스트 데이터셋을 통해 분류를 수행한 결과를 나타낸다.

Table 6. Classification Result of Test Dataset

Activity	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Climbing	5	5	30	33	5	8	33	32
Crawling	8	8	36	23	8	6	36	36
Running	8	6	30	36	8	8	35	37
Walking	3	9	10	34	3	3	32	34
Jumping	13	9	8	29	14	16	22	26

위 Table에서 (a)~(h)의 구분은 Table 4의 내용과 같다. 사용된 테스트 데이터의 개수는 행동별로 40개씩 사용되었으며, 그 중에서 정확하게 분류한 개수를 나타내었다.

분류 결과를 살펴보면 walking, jumping 행동에 대하여 비교적 낮은 분류 결과를 보여주었다. 그러나 전반적인 분류 결과는 Table 4의 검증 데이터셋을 이용한 분류 결과와 비슷한 양상을 보이며, 모든 행동 분류 결과에서 (a), (b), (e), (f)보다 (c), (d), (g), (h)가 분류 성능이 좋았다. Table 7은 분류 결과에 대한 정확도를 나타낸 것이다.

Table 7. Classification Accuracy of Test Dataset

Activity	CNN	LRCN
Climbing	14.38%	80.00%
Crawling	18.75%	81.88%
Running	18.75%	86.25%
Walking	11.25%	68.75%
Jumping	32.50%	53.13%
Total	19.13%	74.00%

테스트 결과, climbing, crawling, 그리고 running은 LRCN을 기준으로 평균 82.70%의 비교적 높은 수치의 정확도를 보인 반면, walking과 jumping의 정확도는 평균 60.93%로 기대에 미치지 못했다. walking과 jumping은 running으로 분류되는 경우가 많았는데, 이들이 running과 유사한 특징을 많이 갖기 때문이라고 판단되며, 클래스 개수 및 학습 데이터의 다양성 부족 때문으로 추정된다. 또한, crawling을 제외한 나머지 행동에 대한 분류 정확도가 validation 정확도에 비해 전부 낮아지는 경향을 보였다. 이는 crawling의 validation set와 test set의 유사성이 다른 행동 부류보다 더 높았기 때문으로 판단된다.

5. 결 론

본 논문은 소규모 행동 클래스를 갖는 분류 문제에서 행동 인식을 위해 기존의 신경망을 최적화하고 실험・평가하였다. 제안한 신경망의 최종적인 분류 정확도는 74.00%로써, 기존 LRCN의 분류 정확도(82.66%)보다 낮다. 이는 다음과 같은 두 가지 관점에서 설명될 수 있다.

첫 번째, 행동 부류의 개수가 기존 LRCN 보다 훨씬 적었다. LRCN은 101가지의 행동을 분류하는 반면, 본 논문의 신경망은 5가지 행동을 분류한다. 기계의 관점에서는 적은 행동 부류의 개수가 오히려 다양한 특징점 확보에 제한사항이 되는 것으로 판단된다.

두 번째, 학습에 사용된 데이터가 기존 LRCN에 비하여 충분치 못하였다. 본 논문에서는 유튜브 검색을 통해 학습 데이터를 확보하였으나, 5가지 침입행동과 관련된 영상 확보가 매우 어려웠다. 학습데이터의 질적・양적인 차이로 인해 분류 정확도가 저하되었을 것으로 판단된다.

그러나 jumping, walking을 제외한 나머지 행동 분류에서는 82.70%의 정확도를 보였으며, 상기한 제한사항을 고려했을 때, 개선을 통해 다른 행동 분류에서도 더 높은 정확도를 얻을 수 있을 것으로 기대된다.

향후 연구과제로 제안된 신경망의 실제 적용 가능성을 검토해보고자 한다. 이를 위해 더미(dummy) 클래스 추가 및 학습데이터 개수를 증가시켜 정확도 향상 여부를 분석해보고, 문제점 개선을 위한 지속적인 연구가 필요할 것이다.

References

[1] E. Kim, S. Helal, and D. Cook, “Human activity recognition and pattern discovery,” *IEEE Pervasive Computing*, Vol.9, No.1, pp.48-53, 2010.

[2] L. Yann, B. Yoshua, and H. Geoffrey, “Deep learning,” *Nature*, Vol.521, No.7553, pp.436-444, 2015.

[3] I. J. Kim, “Recent advances in deep learning technologies for visual recognition,” *Communications of the Korean Institute of Information Scientists and Engineers*, Vol.33, No.9, pp.15-20, 2015.

[4] D. Jeffrey et al, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston-Massachusetts: US, pp.2625-2634, Jun., 2015.

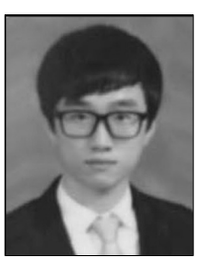
[5] NUREG-1959, Intrusion Detection Systems and Subsystems, USNRC (United States Nuclear Regulatory Commission), p.19, Nov., 2010.

[6] Wikipedia, Activity Recognition [Internet], https://en.wikipedia.org/wiki/Activity_recognition.

[7] C. Liming and K. Ismail, “Activity recognition: Approaches, practices and trends,” in *Activity Recognition in Pervasive Intelligent Environments*, 1st ed. Atlantis Press Pub., ch.3, pp.1-31, 2011.

[8] V. Michalis, N. Christophoros, and I. A. Kakadiaris, “A Review of Human Activity Recognition Methods,” Frontiers in Robotics and AI 2:28, 2015.

[9] J. Kim, C. J. Nan, and B. T. Zhang, “Deep Learning-based Video Analysis Techniques,” *Communications of the Korean Institute of Information Scientists and Engineers*, Vol.33, No.9, pp.21-31, 2015.



김 승 현

e-mail : skim@sunchon.ac.kr

2014년 순천대학교 컴퓨터공학과(공학사)

2016년 순천대학교 컴퓨터과학과(석사)

2016년～현 재 순천대학교 컴퓨터비전 및 보안실험실 연구원

관심분야 : 기계학습, 빅데이터

김 연 호

e-mail : kimyh7102@sunchon.ac.kr
2016년 순천대학교 컴퓨터공학과(공학사)
2016년~현 재 순천대학교 컴퓨터과학과 석사과정
관심분야 : 영상처리, 기계학습

김 도 연

e-mail : dykim@sunchon.ac.kr
1986년 충남대학교 계산통계학과(학사)
2000년 충남대학교 정보통신공학과 (공학석사)
2003년 충남대학교 컴퓨터공학과 (공학박사)
1986년~1996 한국원자력연구원 선임연구원
1997년~2008 한국전력기술(주) 책임연구원
2008년~현 재 순천대학교 컴퓨터공학과 교수
관심분야 : 컴퓨터 비전, 컴퓨터 보안