특집논문 (Special Paper)
방송공학회논문지 제23권 제2호, 2018년 3월 (JBE Vol. 23, No. 2, March 2018)
https://doi.org/10.5909/JBE.2018.23.2.246
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

웨어러블 응용을 위한 CNN 기반 손 제스처 인식

문 현 철[a)], 양 안 나[a)], 김 재 곤[a)‡]

CNN-Based Hand Gesture Recognition for Wearable Applications

Hyeon-Chul Moon[a)], Anna Yang[a)], and Jae-Gon Kim[a)‡]

요 약

제스처는 스마트 글라스 등 웨어러블 기기의 NUI(Natural User Interface)로 주목받고 있다. 최근 MPEG에서는 IoT(Internet of Things) 및 웨어러블 환경에서의 효율적인 미디어 소비를 지원하기 위한 IoMT(Internet of Media Things) 표준화를 진행하고 있다. IoMT에서는 손 제스처 검출과 인식이 별도의 기기에서 수행되는 것을 가정하고 이들 모듈간의 인터페이스 규격을 제공하고 있다. 한편, 최근 인식률 개선을 위하여 딥러닝 기반의 손 제스처 인식 기법 또한 활발히 연구되고 있다. 본 논문에서는 IoMT의 유스 케이스(use case)의 하나인 웨어러블 기기에서의 미디어 소비 등 다양한 응용을 위하여 CNN(Convolutional Neural Network) 기반의 손 제스처 인식 기법을 제시한다. 제시된 기법은 스마트 글래스로 획득한 스테레오 비디오로부터 구한 깊이(depth) 정보와 색 정보를 이용하여 손 윤곽선을 검출하고, 검출된 손 윤곽선 영상을 데이터 셋으로 구성하여 CNN을 학습한 후, 이를 바탕으로 입력 손 윤곽선 영상의 제스처를 인식한다. 실험결과 제안기법은 95%의 손 제스처 인식율를 얻을 수 있음을 확인하였다.

Abstract

Hand gestures are attracting attention as a NUI (Natural User Interface) of wearable devices such as smart glasses. Recently, to support efficient media consumption in IoT (Internet of Things) and wearable environments, the standardization of IoMT (Internet of Media Things) is in the progress in MPEG. In IoMT, it is assumed that hand gesture detection and recognition are performed on a separate device, and thus provides an interoperable interface between these modules. Meanwhile, deep learning based hand gesture recognition techniques have been recently actively studied to improve the recognition performance. In this paper, we propose a method of hand gesture recognition based on CNN (Convolutional Neural Network) for various applications such as media consumption in wearable devices which is one of the use cases of IoMT. The proposed method detects hand contour from stereo images acquisitioned by smart glasses using depth information and color information, constructs data sets to learn CNN, and then recognizes gestures from input hand contour images. Experimental results show that the proposed method achieves the average 95% hand gesture recognition rate.

Keyword : MPEG-IoMT, Hand Gesture, Hand Contour, CNN, Gesture Recognition

a) 한국항공대학교 항공전자정보공학부(Korea Aerospace University, School of Electronics and Information Engineering)
‡ Corresponding Author : 김재곤(Jae-Gon Kim)
E-mail: jgkim@kau.ac.kr
Tel: +82-2-300-0414
ORCID: http://orcid.org/0000-0003-3686-4786
※ 본 연구는 산업통상자원부 국가표준기술원에서 시행한 국가표준기술력향상사업[10077958]의 지원을 받아 수행된 연구임.
※ 이 논문의 연구결과 중 일부는 “한국방송·미디어공학회 2017년 추계학술대회”에서 발표한 바 있음.
※ This work was supported by National Standards Technology Promotion Program of Korean Agency for Technology and Standards (KATS) grant funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (10077958).
※ Parts of this work have been published in the 2017 Fall Conference of the Korean Institute of Broadcasting and Media Engineers.
· Manuscript received January 12, 2018; Revised February 5, 2018; Accepted February 14, 2018.

Ⅰ. 서 론

손 제스처는 스마트 글라스 등의 웨어러블 기기의 NUI (Natural User Interface)로 주목받고 있다. 최근 MPEG에서는 IoT(Internet of Thing) 및 웨어러블 환경에서의 효율적인 미디어 소비를 위하여 IoMT(Internet of Media Things) 표준화를 진행하고 있으며, 손 제스처를 사용한 웨어러블 기기에서의 미디어 소비도 중요한 유스 케이스(use case)로 고려되고 있다. 이와 같이 손 제스처를 웨어러블 기기의 NUI로 사용하기 위해서는 손 제스처의 효율적인 검출 및 인식 기능이 요구된다[1][2][3]. 손 제스처 검출 및 인식을 위한 많은 연구가 진행되었으며[4][5], 최근 딥러닝(Deep Learning) 기술의 발전으로 인하여 다양한 영상 인식 분야에서 딥러닝 알고리즘을 적용하고 있으며 손 제스처 인식을 위한 딥 러닝 연구결과도 보고되고 있다[5][6].

본 논문에서는 MPEG IoMT의 유스 케이스의 하나인 스마트 글래스에서의 미디어 소비를 위한하여 딥러닝의 대표적인 알고리즘인 CNN(Convolutional Neural Network) 기반의 손 제스처 인식 기법을 제시한다. MPEG IoMT에서는 손 제스처 검출과 인식이 별도의 기기에서 수행되는 것을 일반적인 경우로 가정하고, 이들 모듈간의 상호 연동 가능한 API(Application Programming Interface) 및 데이터 포맷 등의 인터페이스를 마련하고 있다[5]. 제안 기법은 영상처리 기반의 손 윤곽선 검출 파트와 CNN 기반의 손 제스처 인식 파트로 구성된다. 손 제스처 검출 모듈에서는 스마트 글래스의 입력 스테레오 영상으로부터 획득한 깊이 정보와 색 정보를 이용하여 손 윤곽선을 이진 영상으로 검출한다. 인식 모듈에서는 검출 모듈에서 얻어진 제스처 이진 영상을 CNN 학습을 위한 데이터 셋으로 사용하고 학습 결과를 바탕으로 손 제스처를 인식한다[6]. 그리고 그 인식결과는 웨어러블 기기로 전송된다.

본 논문에서는 제 2장에서 IoMT의 유스 케이스(use case)의 하나인 제스처 기반 웨어러블 응용 시나리오를 기술하고, 제 3장에서는 제스처 검출 기법 및 검출된 손 윤곽선을 입력하여 제스처 인식을 위한 CNN 기법을 기술한다. 제 4장에서는 제안 기법의 실험결과를 제시하고, 마지막으로 제 5장에서는 본 논문의 결론을 맺는다.

Ⅱ. 제스처 기반 웨어러블 응용 시나리오

IoMT에서는 그림 1과 같은 제스처 기반의 스마트 글라스 응용을 유즈 케이스의 하나로 고려하고 있다[2][3]. 스마트 글라스에서 손 제스처는 양손을 작업 등에 자유롭게 사용할 수 있도록 해주는 중요한 NUI로 부각되고 있다[8]. IoMT는 그림 1과 같이 사용자(User), MThings(Media Things), 처리기 모듈(Processing Unit: PU)으로 구성되며, 사용자는 손 모양과 움직임 경로의 손 제스처를 이용하여 웨어러블 기기를 제어하고 웨어러블 기기를 통하여 다양한 미디어 소비한다.

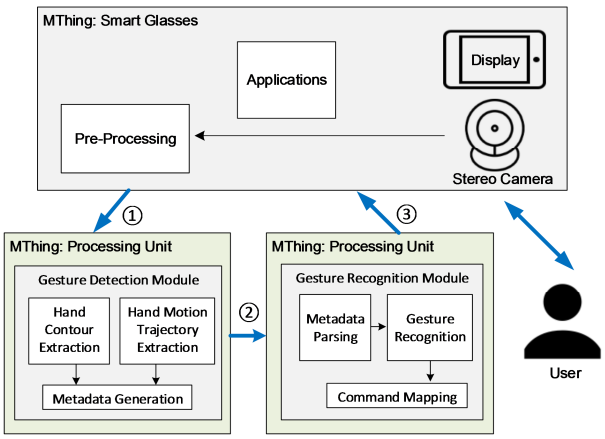

그림 1. 손 제스처 기반의 웨어러블 응용 시나리오
Fig 1. A scenario hand gesture based wearable applications

그림 1과 같이 사용자의 제스처를 포함한 비디오는 스마트 글라스에 내장된 스테레오 카메라를 통해 획득한다. 획득한 제스처 비디오는 전처리 과정을 통해 검출 모듈(Detection Module)로 전달한다. 일반으로 웨어러블 기기의 컴퓨팅 성능이 제한적이기 때문에 웨어러블 환경에서는 검출과 인식이 분리되어 별도의 모듈에서 처리되는 것으로 가정한다[1]. 즉, 검출은 비교적 간단한 연산으로 가능한 반면에, 딥러닝 기반의 인식은 많은 계산량이 필요하다. 따라서 인식 처리는 웨어러블 기기와 연동된 외부 기기에서 처리되는 경우를 가정한다. IoMT에서는 이러한 웨어러블 환경에서 검출과 인식이 별도의 PU에서 처리하기 위해서 PU 간의 상호연동 가능한 인터페이스를 위한 API와 데이터 포

맷을 지원한다. 본 유스 케이스에서는 검출된 손 윤곽선을 표준 포맷에 따라 XML로 기술하고 이를 표준 API를 통하여 전달하고, 인식모듈에서는 검출된 손 윤곽선을 학습된 CNN으로 인식하게 된다.

III. 손 제스처 검출 및 CNN 기반 인식

1. 손 제스처 검출

손 제스처 검출을 위해서 스마트 글라스에 장착된 스테레오 카메라를 이용하여 스테레오 비디오를 획득하고 이로부터 손의 윤곽선을 찾는다. 그림 2과 같이 스테레오 카메라로 입력된 좌, 우 영상(RGB 영상)의 차이를 스테레오 매칭을 통하여 깊이(depth) 영상을 획득한다. 스마트 글래스 특성상 사용자의 손은 카메라로부터 일정한 거리 범위(30 ~ 50cm)에 있다고 가정을 하고, 이를 바탕으로 깊이 영상과 색정보를 적용하여 손 제스처 영역을 검출한다. 보다 정확한 손 제스처의 영역을 얻어내고 잡음을 제거하기 위해 모폴로지(morphology) 연산을 한다. 그림 2는 본 논문의 손 영역 검출 전 과정을 보여준다. 그림 3은 제시된 기법의 검출 결과 예이다[1][2][3].

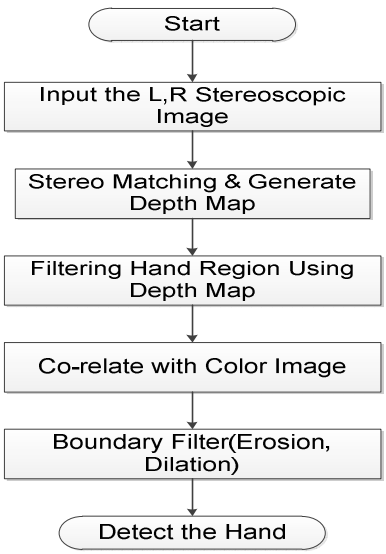

그림 2. 손 윤곽선 검출 순서도
Fig 2. Procedure of hand contour detection

그림 3. 검출된 손 윤곽선 예
Fig 3. An example of the detected hand contour

2. CNN 기반 제스처 인식

CNN은 다층신경망의 한 종류로 인식 분야에서 매우 좋은 결과를 보이고 있다. CNN은 기존의 신경망과 다르게 입력 원본 데이터를 바로 연산하여 출력하지 않고, 특징(feature) 추출 단계와 분류화(classification) 단계를 거쳐 결과값을 내는 것이 특징이다. CNN의 구조는 컨벌루션(convolution) 층, 풀링(Pooling) 층, 완전연결(Fully-Connected: FC) 층으로 구성되어 있다[9].

그림 4는 본 논문의 제스처 검출을 위한 CNN 구조도이다. C1, C2는 Convolution 층으로 입력계층에서 입력되는 영상 데이터를 5×5 크기의 필터를 사용하여 간격 1(stride=1)로 컨벌루션 연산을 수행하고, 입력 손 윤곽선 영상의 특징을 추출해 낸다. S1, S2는 Pooling 층으로 여기서는 Max Pooling 기법을 사용하여 특징 맵의 크기를 2×2, 즉 가로와 세로크기를 각각 반으로 줄이는 역할을 수행한다. Max Pooling 층은 설정한 크기 범위 내에서 가장 특징이 두드러진 부분을 그 범위의 대표 값으로 지정하는 방식으로 그림 5는 그 예를 보여주며, 설정된 크기 범위는 2×2이다.

F1, F2는 Fully-Connected 층이며, 이 계층은 일반적인 다층신경망의 구조와 같이 추출된 특징을 바탕으로 신경망에서 연산하여 출력 값을 결정한다. 그림 6은 Fully-Connected층의 분류화 과정의 간단한 예를 나타낸 것이며, 여기서 P1, P2, P3는 영상에서 추출된 각 클래스에 해당하는 확률 값을 나타낸다. 그림 6의 경우는 예시이며, 본 논문에서는 그림 7의 제스처를 인식하고자 하므로, 각 제스처의 확률 값 P1~P10로 변경된다. 확률 값이 계산되면 Softmax

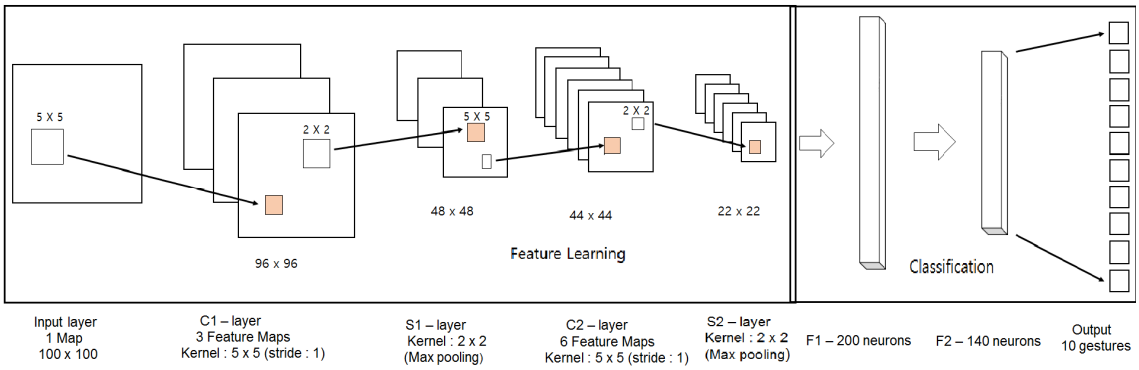

그림 4. 제안하는 CNN 구조
Fig 4. Proposed CNN structure

함수를 사용하여 확률 값이 가장 큰 제스처가 입력 영상의 제스처로 인식하고,(결과 값 '1') 나머지 9개의 제스처는 결과 값 '0'으로 된다.

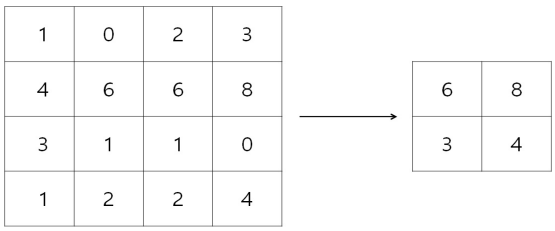

그림 5. Max Pooling 예
Fig 5. Example of Max Pooling

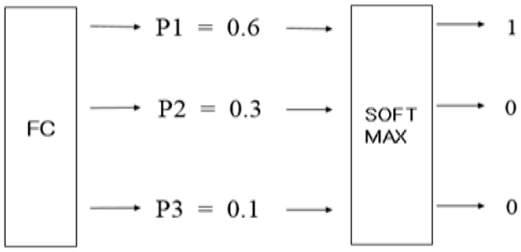

그림 6. Fully-Connected 층의 분류화 과정
Fig 6. Classification process of fully-connected layer

3. CNN 최적화 알고리즘

CNN 학습과정에서 비용함수를 최소화하도록 각 층의 가중치(weight)를 학습시키는데, 가중치는 입력 값에 곱해져 출력 값을 내게 된다. 이때 가중치를 조절하기 위해서 최적화 알고리즘을 사용한다. 대표적으로 기울기감소법(gradient descent), 모멘텀(momentum) 등이 있다. 기울기 감소법은 네트워크의 출력 결과 값과 실제 값의 오차함수인 손실함수(loss function)를 이용하는데, 가중치를 조절하여 손실함수가 최소 값이 되는 방향으로 이동하게 하는 알고리즘이다. 이 때 모멘텀은 기울기감소법을 통해 이동하는 과정에 과거에 이동했던 결과를 기억하면서 그 방향으로 일정 정도를 추가적으로 이동하는 방식으로 아래의 식으로 표현된다[10].

$$v_{t+1} = \mu v_t - \epsilon \nabla f(\theta_t) \quad (1)$$

$$\theta_{t+1} = \theta_t + v_{t+1} \quad (2)$$

위의 수식 (1), (2)은 모멘텀을 나타낸 것으로 $\nabla f(\theta_t)$는 t에서의 gradient , ϵ는 기울기 감소법에서의 이동 간격인 learning rate, μ는 모멘텀을 얼마나 줄 것인지에 대한 모멘텀에 대한 가중치 계수이며, v_t는 t에서의 이동벡터이다. 수식 (1)에서 v_t를 현재 이동벡터라고 가정한다면, 다음 반복인 v_{t+1}에 현재의 gradient와 현재 이동벡터가 미리 정해진 계수에 곱해져 누적된다. 즉, v_{t+1}은 이전까지 누적된 gradient이다. 이는 수식 (2)에서 다음 위치 θ_{t+1}를 정할 때, 현재의 위치 θ_t에 누적된 그래디언트를 합쳐서 다음 t+1의 방향을 정하게 된다. 본 논문에서는 모멘텀 기법의 적용 여부에 따른 최적화 알고리즘의 성능을 분석하였다.

4. 딥러닝 프레임 워크

딥러닝 프레임워크들은 Tensorflow, Theano, MXnet, Keras 등이 있으며, 예측하거나 분석하려는 대상에 따라 맞는 프레임워크를 선택하면 좋은 성능을 얻어낼 수 있다[11]. 본 논문에서의 S/W 언어는 통계나 병렬계산에 특화되어 있는 R 언어를 사용했다. 따라서 R언어에서 활용할 수 있는 MXnetR, Deepnet, H20 등을 통해 실험했으며, 각각의 프레임워크별로 내부함수 형태가 다르게 설정되어 있다. 예를 들면, MXnet의 경우 심볼릭 프로그래밍과 명령형 프로그래밍의 혼합된 형태를 허용해서 효율성과 유연성을 극대화하기 위해 설계된 프레임워크이며, Deepnet은 다른 프레임워크에 비해 상대적으로 작지만 다양한 구조형식을 선택할 수 있게 설계되어 있다. 또한 H20는 분산 컴퓨터 시스템을 활용할 수 있으며, 예측분석 플랫폼으로 다양한 프로그래밍에 대한 인터페이스를 제공한다. 본 논문에서는 각 프레임워크별 제스처 인식 정확도를 비교하였다.

IV. 실 험 결 과

본 논문의 실험에서는 학습을 위한 데이터 셋으로 8명의 피실험자의 손 제스처 영상을 스테레오 카메라로 획득하였다. 그림 7의 10개의 제스처에 대한 피실험자 5명의 6,000개의 학습 데이터, 3명의 1,000 개의 테스트 데이터를 구성하고 인식의 정확도를 확인하였다.

그림 7. 실험에 사용한 손 제스처
Fig 7. A set of hand gestures used in the experiments

표 1은 딥러닝 프레임워크별 실험결과 정확도를 비교한 것이다. 표에서 정확도는 인식율을 나타낸 것으로, 테스트 데이터 1,000개에 대한 정확도이다. 각 프레임워크에 따라서 내부함수 구현방식이 다르며, 이에 따라 인식의 정확도가 다르다. 실험은 앞에서 제시한 같은 구조로 사용했으며, 표 1의 실험결과를 바탕으로 인식 정확도가 가장 높은 MxnetR를 이후 실험에서 사용하였고, 표 2와 표 3은 MxnetR을 사용한 결과이다.

표 1. 딥러닝 프레임워크 별 인식 정확도
Table 1. Recognition accuracy comparisons among deep learning frameworks

Framework	Accuracy (%)
MxnetR	95
Deepnet	94.7
H20	94.4

표 2는 데이터 셋의 크기와 모멘텀 기법의 사용 여부에 따른 정확도 비교이다. 여기서 데이터 셋의 크기 는 학습 데이터와 테스트 데이터의 합이며, 학습 데이터는 5명의 피실험자에 대한 데이터이고, 테스트 데이터는 3명의 피실험자에 대한 데이터이다. 본 실험에서는 데이터의 수가 많아짐에 따라 정확도가 높아지는 것을 알 수 있으며, 데이터의 수가 7,000개일 때 제일 높은 정확도를 보여준다. 모멘텀의 적용 유무에 따른 0.3%정도의 정확도 차이가 있으며, 모멘텀 유무에 따라 성능의 차이를 확인할 수 있었다.

표 2. 데이터 셋 수와 모멘텀 적용에 따른 인식 정확도
Table 2. Recognition accuracy comparisons according to the size of data set and the application of momentum

Methods	Accuracy (%)
Data set = 2,800 Momentum: not applied	77.2
Data set = 5,600 Momentum: not applied	88.4
Data set = 7,000 Momentum: not applied	94.7
Data set = 7,000 Momentum: applied	95

표 3은 그림 6의 테스트 제스처 별 인식 정확도를 나타낸다. 3을 나타내는 “Three” 제스처의 정확도가 98.3%으로 제일 높으며, 주먹을 나타내는 “Rice” 제스처의 정확도가 91.7%로 제일 낮았다. 각 제스처에 대한 정확도의 차이가 약 0.5 ~ 6.6%까지 보였으며, 10개 제스처 모두의 평균 정확도는 95%이다.

표 3. 각 제스처의 인식 정확도
Table 3. Recognition accuracy of each gesture label

	1 (One)	2 (Two)	3 (Three)	4 (Four)	5 (Five)
Accuracy (%)	93.1	96.1	98.3	95.6	95.2
	6 (Okay)	7 (Promise)	8 (Rice)	9 (Scissor)	10 (Victory)
	94.2	96.1	91.7	93.7	94.6
	Average accuracy = 95 %				

표 4는 CNN을 이용해 이미지를 전처리 과정을 통해 이진화 영상을 데이터 셋으로 사용하고, 이를 CNN을 이용해 인식하는 기존 기법 [12]와 본 논문의 정확도를 비교한 것이다. 기존 기법과 본 논문의 차이점은 전처리 과정에서의 깊이 맵의 이용 여부이며, 이는 배경이 더 복잡할수록 검출의 정확도의 차이가 있다. 논문 [12]의 정확도는 93.8%이며, 본 논문과의 정확도는 1.2% 차이가 난다.

표 4. 기존 기법과의 인식 정확도 비교
Table 4. Comparison of recognition accuracy with the existing method

Method	Accuracy (%)
Existing method [12]	93.8
Proposed method	95

V. 결 론

본 논문에서는 웨어러블 응용을 위한 손 제스처 인식을 기법을 제시하였다. 제시된 기법은 영상처리 기반의 손 제스처 영역 검출과 CNN을 이용한 제스처 인식으로 구성된다. 제스처 검출에서는 스마트 글래스의 스테레오 카메라로 획득한 스테레오 영상의 깊이 정보와 색 정보를 이용하여 손 영역을 검출하였고, 인식에서는 검출된 손 영역 이진 영상을 데이터 셋으로 이용하여 CNN을 학습하고 이를 통해 제스처 인식을 하였다. 학습 데이터 셋의 크기, 최적화 기법, CNN 프레임워크에 따른 인식 정확도 성능을 분석하였다. 실험결과 95%의 높은 인식 정확도 결과를 얻었으며, 제시된 기법은 MPEG IoMT에서의 손 제스처 기반의 웨어러블 응용에 활용할 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] A. Yang, S. Chun, H. Ko, J. G. Kim, “Hand gesture description for wearable applications in M-IoTW,” ISO/IEC JTC1/SC29/WG11 M38526, Geneva, Swiss, May. 2016.
[2] A. Yang, S. Chun, and J-G. Kim, “Detection and recognition of hand gesture for wearable applications in IoMTW,” In Proc. ICACT 2017, pp. 598 - 601, Feb. 2017.
[3] A. Yang, S. Chun, and J.-G. Kim, “Detection and Recognition of Hand Gesture for Wearable Applications in IoMTW,” ICACT Trans. Advanced Communications Technology (TACT), vol. 6, no. 5, pp. 1046-1053, Sep. 2017.
[4] S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Trans. Syst., Man, Cybern. C, vol. 37, no. 3, pp. 311 - 324, 2007.
[5] S. Byun, S. Lee, G. Kim, S. Han, “Gesture recognition with wearable device based on deep learning,” Broadcasting and Media Magazine, Vol.22, No.1, pp. 58-66, Jan.2017
[6] H. Moon, A. Yang, S. Chun, and J-G. Kim, “CNN-based hand gesture recognition for wearable applications,” The Korean Institute of Broadcase and Media Engineers Conference, Seoul, Korea, pp. 58-59, 2017.
[7] Y. LeCun, K. Koray, and F. Clément, “Convolutional networks and applications in vision,” In Proc. ISCAS 2010.
[8] M. Mitrea, “Working draft 2.0 of ISO/IEC 23093-1 IoMT Architecture,” ISO/IEC JTC1/SC29/WG11 N17094, Torino, July 2017.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagent classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, 2012.
[10] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” In Proc, Int, Conf. Machine Learning (ICML), pp 1139-1147, Feb, 2013.
[11] Y. Lee, P. Moon, “A comparison and Analysis of deep learning framework,” J. Korea Institute of Electron. Communi. Science (KIECS), vol 12, no.1, pp. 115-122, Feb, 2017.
[12] M. Han et al., “Visual hand gesture recognition with convolution neural network,” In Proc. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2016.

저 자 소 개

문 현 철

- 2018년 2월 : 한국항공대학교 항공전자정보공학부 학사
- 2018년 3월 ~ 현재 : 한국항공대학교 항공전자정보공학과 석사과정
- ORCID : http://orcid.org/0000-0002-1672-2345
- 주관심분야 : 비디오 부호화, 영상처리, 딥러닝

양 안 나

- 2014년 7월 : 한국항공대학교 항공전자 및 정보통신공학부 학사
- 2017년 2월 : 한국항공대학교 항공전자정보공학과 석사
- 2018년 3월 ~ 현재 : 한국항공대학교 항공전자공학과 박사과정
- ORCID : https://orcid.org/0000-0003-4957-9589
- 주관심분야 : 비디오 부호화, IoT 및 웨어러블 미디어, 영상처리

김 재 곤

- 1990년 2월 : 경북대학교 전자공학과 학사
- 1992년 2월 : KAIST 전기 및 전자공학과 석사
- 1992년 2월 : KAIST 전기 및 전자공학과 박사
- 1992년 3월 ~ 2007년 2월 : 한국전자통신연구원(ETRI) 선임연구원/팀장
- 2001년 9월 ~ 2002년 7월 : 뉴욕 콜롬비아대학교 연구원
- 2015년 12월 ~ 2016년 1월 : UC San Diego, Visiting Scholar
- 2007년 9월 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
- ORCID : http://orcid.org/0000-0003-3686-4786
- 주관심분야 : 비디오 부호화, 비디오 신호처리, 영상통신, UHD/실감미디어