웨어러블 센싱 음향-영상 라이프로그 학습 기반 사용자 활동 인식*

이충연[1○], 곽동현[2], 곽하녹[1] 장병탁[1,2]
[1]서울대학교 컴퓨터공학부, [2]서울대학교 뇌과학협동과정
{cylee, dhkwak, hnkwak, btzhang}@bi.snu.ac.kr

Activity Recognition by Learning Auditory-Visual Lifelogs from Wearable Sensors

Chung-Yeon Lee[1○], Dong-Hyun Kwak[2], Hanock Kwak[1], Byoung-Tak Zhang[1,2]
[1]Department of Computer Science and Engineering, Seoul National University,
[2]Interdisciplinary Programs in Neuroscience, Seoul National University

요 약

최근 차세대 스마트 기기로 주목 받고 있는 웨어러블 디바이스(Wearable devices)는 하나 또는 여러 개의 센서를 탑재하여 사용자가 실제로 접하는 다양한 정보를 보다 자연스럽게 기록하기에 적합하다. 하지만 수시로 수집되는 순수 센서 데이터들은 신호 자체만으로는 그 의미를 명확하게 파악하기가 어렵고 서로 다른 특성을 가지는 이질적인 양상을 보이며, 시간에 따라 변화하는 동적인 특성을 가지기 때문에 각각의 신호 데이터 특성에 맞는 핵심 정보를 추출해내는 전처리 과정이 필요하며, 상황에 따라 실시간으로 변화하는 환경과 사용자 행동 간의 유의미한 연관성을 파악하는 방법이 개발되어져야 한다. 본 논문에서는 웨어러블 디바이스를 이용하여 사용자 일상에서 획득 가능한 일인칭 기준 시점의 영상과 음향 데이터로부터 딥 컨볼루션 신경망과 MFCC 계수를 이용하여 유의미한 특징 벡터를 추출하고, 이를 이용하여 사용자의 활동 내역을 분류하는 방법을 제안한다. 또한 제안하는 방법을 검증하기 위해 스마트폰을 통해 기록되는 Logging App을 만들어 활동 내역과 위치 정보를 함께 수집하여 인식 성능 평가를 수행하였다.

1. 서 론

최근 차세대 스마트 기기로 주목 받고 있는 웨어러블 디바이스(Wearable devices)는 우리 몸의 여러 부위에 착용이 가능한 형태로 설계된 정보 기기들을 통칭한다. 대표적으로 Google이 개발한 스마트 안경 형태의 Google Glass, 스마트 시계인 삼성의 Galaxy Gear, 클립이나 밴드 형태로 몸에 착용 가능한 운동량 측정 기기인 Fitbit과 Misfit Shine 등이 있다. 웨어러블 디바이스는 하나 또는 여러 개의 센서를 탑재하여 사용자의 각종 정보를 일상에서 자연스럽게 획득할 수 있다. 특히 Microsoft가 2014년 출시한 MS Band는 심박수, GPS, 가속도, 피부전도도를 포함하여 총 10개 종류의 센서를 탑재하고 있어 다양한 응용에 사용될 수 있을 것으로 기대된다. 하지만 아직까지 이처럼 다양한 웨어러블 센서 데이터를 함께 사용하여 2차 정보를 생성하거나 이를 통해 응용 서비스를 제공하는 사례는 찾아볼 수 없었다. 웨어러블 디바이스가 단순히 스마트폰의 일부 기능을 대체하는 수준에서 그치지 않고, 사용자에게 특화된 경험을 제공하는 서비스로 포지셔닝하기 위해서는 다양한 센서 정보를 융합하여 차별화된 서비스를 제공할 수 있어야 할 것이다. 단, 수시로 수집되는 순수 센서 데이터들은 신호 자체만으로는 그 의미를 명확하게 파악하기가 어렵고, 서로 다른 특성을 가지는 이질적(Heterogeneous)인 양상을 보이며, 시간에 따라 변화하는 동적인 특성을 가진다. 따라서 각각의 신호 데이터 특성에 맞는 핵심 정보를 추출해내는 전처리 과정이 필요하고, 또한 상황에 따라 실시간으로 변화하는 환경과 사용자 행동 간의 유의미한 연관성을 파악하는 방법이 개발되어져야 한다[1].

본 논문에서는 웨어러블 디바이스를 이용하여 사용자 일상에서 획득 가능한 일인칭 기준 시점(Egocentric)의 영상과 음향 데이터로부터 유의미한 핵심 정보를 추출하고, 이를 이용하여 사용자의 활동 내역을 분류하는 방법을 제안한다. 이를 위해 다양한 영상 인식 분야에 적용되어 매우 높은 성능을 보이고 있는 딥 컨볼루션 신경망(Deep convolutional neural networks, CNN)[2]과 사람의 소리 인지 주파수를 반영하는 것으로 알려져 있는 MFCC (Mel-frequency Cepstral Coefficients)를 각각 영상과 음향의 특징 벡터를 추출하는데 사용하였다.

또한 제안하는 방법을 검증하기 위해 스마트폰을 통해 사용자가 임의로 입력 가능한 Activity Logging App을 만들어 식사, 수업, 휴식 등 실험 참가자의 활동 내역과 위치 정보를 함께 수집하고 인식 성능 평가를 수행하였다.

* 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(R0126-15-1072-SW스타랩, 10035348-mLife, 10044009-HRI.MESSI)과 한국연구재단의 지원(NRF-2010-0017734-Videome), 네이버(NAVER)의 지원을 받아 수행된 연구임.

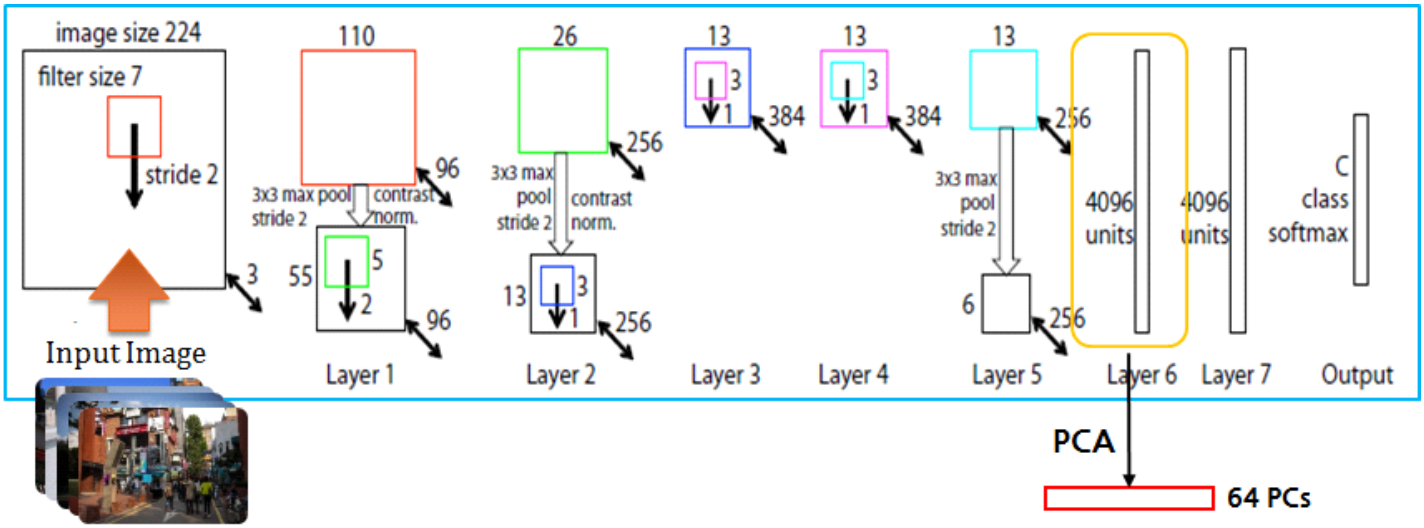

그림 1. 영상 특징 벡터 추출을 위한 딥 컨볼루션 신경망 구조

2. 연구 내용 및 방법

2.1 영상 특징 벡터 추출 방법

Google Glass와 같은 스마트 안경을 이용하여 녹화된 일인칭시점 비디오로부터 1초에 한 장씩의 이미지를 추출한 후, 1000개 클래스의 ImageNet Database (ILSVRC) [3]를 학습한 딥 컨볼루션 신경망(Deep convolutional neural networks, CNN)을 이용하여 각 이미지로부터 확률적 특징 벡터를 추출한다. 이때 계산에 소요되는 시간을 단축하기 위해 CUDA GPGPU 병렬 처리 컴퓨팅이 가능한 MatConvNet을 사용한다 [4].

그림 1은 특징 벡터 추출에 사용된 CNN 구조이며, 대체로 이미지 인식을 위해 Layer 8까지가 보편적으로 사용된다. 특징 벡터의 추출 과정은 크게 컨볼루션과 Max-pooling 단계로 이루어진다. 먼저 컨볼루션 레이어의 특징맵 노드들은 각각 하위 레이어의 특정 위치 윈도우 내 노드들과 연결된다. 이때 동일한 특징맵의 노드들은 동일한 가중치를 공유하며, 이는 기존의 영상 처리 기법에서 사용되는 컨볼루션과 동일한 효과를 보인다. 여기에서 가중치는 컨볼루션 마스크의 역할을 수행하는데, 기존 컨볼루션 알고리즘이 고정된 마스크를 이용한 반면 CNN에서는 컨볼루션 마스크가 신경망으로 구현되기 때문에 데이터로부터 자동으로 학습된다는 차이가 있다.

Max-pooling 레이어는 연결된 컨볼루션 레이어와 동일한 수의 특징맵을 가지며 입력 특징맵과 1:1로 연결된다. 각 노드들은 자신과 연결된 입력 노드들의 값 중 최대값을 선택하여 가져오는 방식으로 동작한다. 이러한 방식은 윈도우 내 좌표는 무시하기 때문에 각 특징이 추출되는 위치상의 변이에 영향을 받지 않아, 결과적으로 기하학적 변화에 강인한 특성을 가지게 된다 [5].

최근 연구 결과에 따르면 이렇게 학습된 CNN을 특징 추출을 위해 사용하여 전이학습(Transfer learning)이 가능하다 [6]. 본 논문에서는 CNN의 6번째 레이어의 4096차원 값을 주성분분석(PCA)을 이용하여 64차원으로 축소하고, 이를 영상 데이터의 특징 벡터로 사용한다.

2.2 음향 특징 벡터 추출 방법

위와 동일한 비디오 데이터에서 음향 데이터를 분리한 후, 사람의 소리 인지 주파수를 반영하는 것으로 알려져 있는 MFCC 계수(Mel-frequency cepstral coefficient)를 그림 2와 같이 추출한다. 이때 8000 Hz의 샘플링 레이트로 기록된 음향 데이터를 2000개 샘플(250 ms)로 구성된 프레임을 이전 프레임과 50% 중첩되는 방식으로 Sliding windowing하면서 FFT (Fast Fourier transform)를 이용하여 주파수 영역으로 변환시킨 후, Mel-scale의 필터 뱅크(Filter bank)에 통과시켜 파워스펙트럼을 구한다. 이후 Mel-scale 필터 뱅크를 통과한 데이터에 로그 함수와 DCT (Discrete cosine transform)를 수행하여 12개의 주파수 특성과 하나의 프레임 로그 에너지 특성으로 구성된 총 13차의 MFCC 계수를 추출한다. 최종적으로 음향 데이터로부터 1초마다 7개의 13차원 특징 벡터들이 구해지게 되며, 분류 단계에서는 2차원 특징 벡터를 91개 값으로 구성된 1차원 특징 벡터로 변환하여 사용한다.

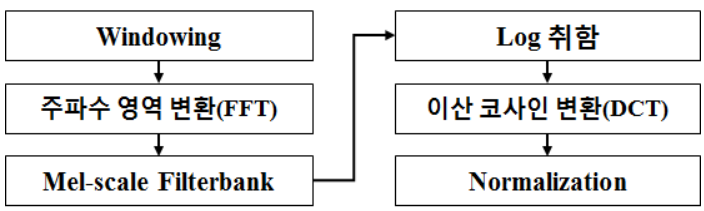

그림 2. MFCC 계수 추출 방법

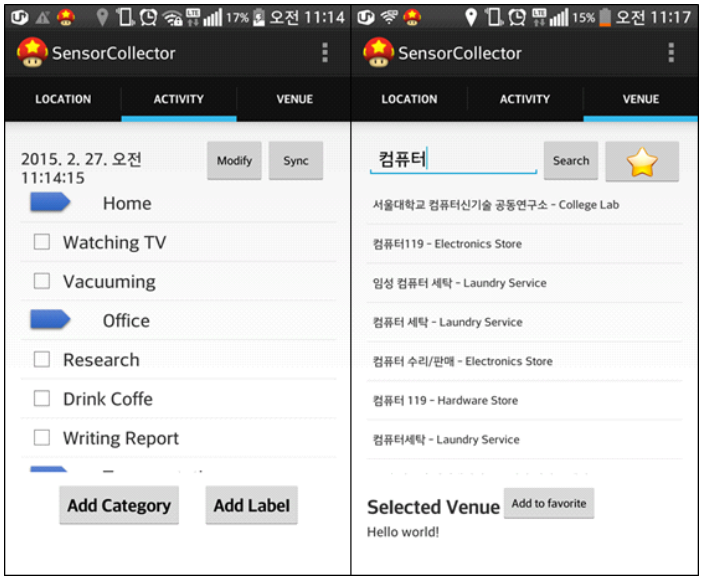

그림 3. 사용자 활동 내역 Logging App

3. 실험 결과 및 논의

본 실험을 위해 2명의 실험 참가자가 각각 14일 동안 스마트 안경을 착용하고 생활하면서 약 150시간 분량의 일인칭 비디오 데이터를 수집하였으며, 같은 기간 동안 그림 3과 같은 스마트폰 Activity Logging App을 이용하여 사용자의 위치와 활동 내역을 기록하였다. 표 2는 일련의 전처리 과정을 통해 정리된 사용자의 활동 내역과 각 활동별 데이터 길이를 나타낸 것이다.

제안한 방법으로 추출한 영상과 음향의 특징 벡터를 서로 연결하고 K-nearest neighbors classifier (KNN)를 이용하여 5-fold cross-validation으로 분류한 결과 82.85%의 인식률을 보였다. 그림 4는 분류 결과에 대한 Confusion matrix이다. 통제되지 않은 실생활에서 서로 다른 참가자가 획득한 데이터들과 구분하기에 다소 모호한 활동 내역을 레이블로 사용한 실험이기 때문에 이러한 분류 결과는 매우 고무적이라고 볼 수 있다.

4. 결 론

본 논문에서는 웨어러블 디바이스에 장착된 멀티모달 센서를 통해 사용자의 일상에서 자연스럽게 획득 가능한 일인칭 기준 시점의 영상과 음성 데이터로부터 딥 컨볼루션 신경망과 MFCC 계수를 이용하여 유의미한 특징 벡터를 추출하고, 이를 이용하여 사용자의 활동 내역을 분류하는 방법을 제안하였다. Activity Logging App을 통해 함께 수집된 실험 참가자의 12가지 활동 내역을 분류한 실험 결과는 높은 인식률을 보여, 웨어러블 센싱 데이터가 사용자의 활동 내역을 잘 반영하고 있음이 검증되었다.

표 2. 사용자 활동 내역 및 데이터 수집 시간 (단위: 초)

Meal	Lecture	Meeting	Research	Rest	Work
68,206	15,469	6,398	81,327	127,837	51,220
Bike	**Bus**	**Car**	**Subway**	**Walk**	**Shopping**
2,305	36,569	16,848	18,537	89,737	2,265

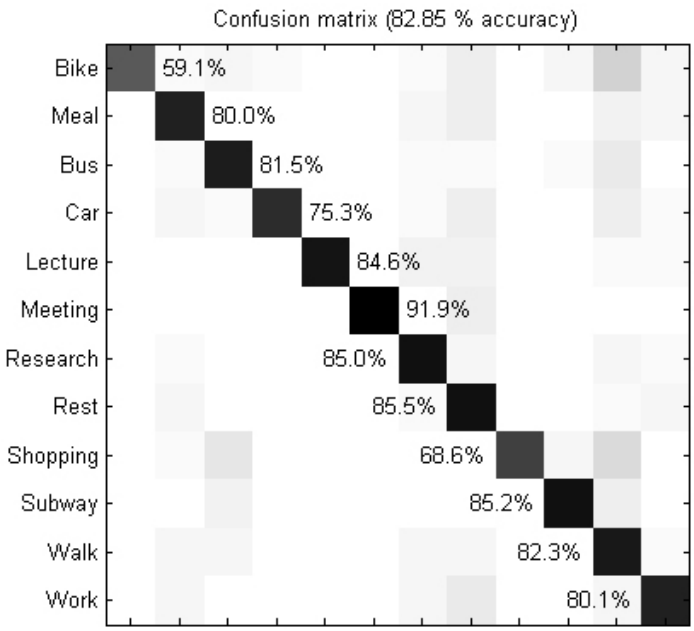

그림 4. 사용자 활동내역 분류 결과

참 고 문 헌

[1] 이충연, 이범진, 온경운, 하정우, 김홍일, 장병탁, "모바일 멀티모달 센서 정보의 앙상블 학습을 이용한 장소인식," *정보과학회 컴퓨팅의 실제 논문지*, 제 21권, 제 1호, pp. 64—69, 2015.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems 25 (NIPS 2012)*, pp. 1097—1105, 2012.

[3] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," arXiv:1409.0575, 2014.

[4] A. Vedaldi, K. Lenc, "MatConvNet—convolutional neural networks for MATLAB," arXiv:1412.4564, 2014.

[5] 김인중, "Deep Learning: 기계학습의 새로운 트렌드," *한국통신학회지(정보와통신)*, 제 31권, 제 11호, pp. 52—57, 2014.

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 3320—3328, 2014.