사용자 활동 인식을 위한 다중 웨어러블 데이터 분석 기법

온경운[O1], 김은솔[1], 장병탁[1,2,3]
[1]서울대학교 컴퓨터공학부, [2]인지과학 협동과정, [3]뇌과학 협동과정
{kwon, eskim, btzhang}@bi.snu.ac.kr

Analyzing multi-wearable data for recognizing user activity

Kyoung-Woon On[O1], Eun-Sol Kim[1], Byoung-Tak Zhang[1,2,3]
[1]School of Computer Sci. & Eng., [2]Brain Science Program, [3]Cognitive Science Program, Seoul National University

요 약

본 논문은 사람이 다중 신호 정보를 처리하는 방식을 모방하여 웨어러블 센서로부터 수집된 다중 센서 데이터를 통합하여 사용자의 활동을 예측하는 모델을 제안한다. 피험자는 2주동안 구글 글래스를 착용하고 일상적인 활동을 하면서 1인칭 시점의 비디오, 오디오 데이터를 수집하였다. 실제 생활 환경에서 수집한 데이터는 제약조건이 없고 잡음이 많은 특성이 있는데 본 논문에서는 이러한 데이터를 효율적으로 분석할 수 있는 기계학습 모델을 제시한다. 제안하는 모델은 베이지안 네트워크를 이용하여 센서 데이터의 근원을 파악, 같은 근원의 데이터는 통합하고 그렇지 않은 경우엔 분리하여 처리한다. 실험 결과로 제안하는 알고리즘을 통하여 피험자의 활동(activity)를 추론하는 결과를 확인하였다.

1. 서 론

최근 들어 사용자의 몸에 부착하여 행동 데이터를 수집할 수 있는 웨어러블 장치가 많이 등장함에 따라 컴퓨터 공학 및 인지과학 분야 연구에 큰 변화가 생기고 있다. 웨어러블 장치는 몸에 부착하는 형식으로 피험자의 행동을 방해하지 않고 여러 가지 반응 데이터를 분석할 수 있기 때문이다.[1] 예를 들어, 신체에 부착된 가속도계 센서는 몸의 움직임이나 활동(activity)을 분석하는데 사용된다.[2] 또한, 안경형 시선 추적기(Eye-tracker)에서 추출되는 시선 정보는 사람의 주의, 흥미에 대한 분석에 사용된다.[3] 또한 구글 글래스에서 사용자 1인칭 시점의 비디오 정보와 그 당시의 오디오 정보를 이용하여 사용자의 현재 활동에 대한 분석을 할 수 있다. 이러한 웨어러블 센서 데이터는 사람의 정보 처리 메커니즘을 연구하고, 나아가 사람을 닮은 인공지능 시스템을 연구하는데 사용될 수 있다.

하지만 센서 데이터는 다양한 외부 환경 변화와 내부적 요소로 인해 부정확하고 잡음이 많은 정보를 제공한다. 예를 들면, 피부 온도를 측정하는 웨어러블 장비는 외부 기온의 변화에 크게 영향을 받을 수 있다. 부정확한 데이터는 해당 센서 정보로부터 올바른 추론을 하는데 걸림돌이 된다. 이러한 센서 데이터의 잡음은 다른 종류의 센서 데이터와 통합함으로써 그 효과를 감소시킬 수 있다.[4] 하지만 각 센서 데이터가 가지고 있는 정보량은 서로 다르고, 상황에 따라 주요한 센서와 그렇지 않은 센서가 달라지기 때문에 동등하게 통합하는 것은 문제가 될 수 있다.

따라서 본 논문에서는 복합 센서 데이터를 통합하고 데이터로부터 의미 있는 정보를 추출할 수 있는 새로운 앙상블 모델을 제안한다. 해당 모델은 사람의 센서 정보 처리 방식을 모방할 수 있는 베이지안 신호 결합 모델을 기반으로 한다.[4] 다음 장에서는 모델을 설명하고 그 후 사용한 데이터, 실험 결과를 소개하고 결론에 대해 이야기한다.

2. 모 델

본 연구에서는 다양한 종류의 센서 정보를 효율적으로 처리할 수 있는 모델을 제안한다. 해당 모델은 사람이 다중 신호 정보를 처리하는 방식을 모방하였다. 사람의 뇌는 외부로부터 들어오는 신호 정보를 처리할 때 신호의 근원을 이해하여 정확한 인지활동을 수행하는데 이를 위해 신경계에서는 단순히 여러 신호 정보를 통합하지 않고, 신호가 같은 근원으로부터 발생되었을 때 통합하고 그렇지 않을 땐 개별적으로 처리한다.[4] 따라서 제안하는 모델은 이러한 관계를 베이지안 네트워크로 모델링하여 각 센서 데이터 처리 모듈을 근원에 대한 정보에 따라 확률적 방법으로 통합한다.

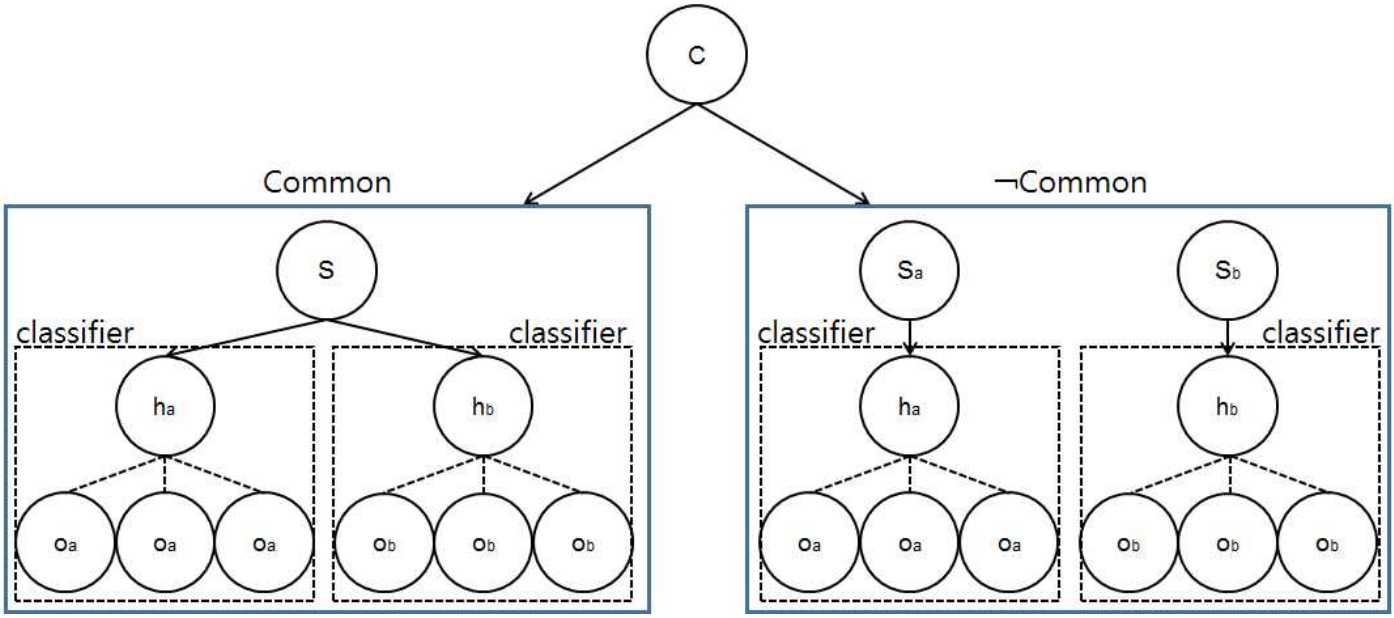
전체 모델의 구조도는 [그림 1]과 같다. 센서 a, b로부터 입력되는 센서 정보는 o_a, o_b이고 각 센서를 처리하는 모듈로부터 각각 더 추상적인 정보인 h_a, h_b를 얻을 수 있다. 본 연구는 초기 실험으로서 o_a, o_b를 쉽게 결합하기 위해 각 처리 모듈을 Naive Bayes 분류기로 설정하여 같은 domain을 가지는 h_a, h_b로 변환하였다. 따라서 각 센서 처리 모듈은 단일 센서 데이터 a, b를 이용하여 학습시킨 분류기이고 h는 결과(클래스 레이블)를 담고 있는 변수이다. 또한 변수 S는 실제 데이터의 클래스 레이블을 나타낸다. 이해를 돕기 위해 실제 실험에 사용된 데이터를 예로 들면, 클래스 레이블은 피험자의 특정 활동(activity)를 나타내고 특정 활동을 수행할 때 구글 글래스의 1인칭 시점 비디오, 오디오 정보는 각각 o_a, o_b가 된다. 따라서 특정 활동(클래스 레이블)은 센서 데이터인 비디오 오디오 정보의 근원으로 설명할 수 있다. 그리하여 해당 모델에서의 클래스 레이블은 센서 정보의 근원을 나타내고, 같은 근원일 경우(Common)와 다른 근원일 경우(⊟Common)로 나누어 같은 근원일 경우 하나의 클래스 레이블 S를 추정하고, 그렇지 않은 경우에는 각각의 추정치 S_a, S_b를 계산한다. 이 때 $p(S)$, $p(H_a|S)$, $p(H_b|S)$는 식(1), (2), (3)과 같이 categorical 분포를 따른다.

$$P(S) = Cat(K, \boldsymbol{p_s}) = \prod_{i_s=1}^{K} p_{i_s}^{[s=i_s]} \quad (1)$$

$$P(H_a|S) = Cat(K, \boldsymbol{p_{H_a|S}}) = \prod_{i_{H_a|S}=1}^{K} p_{i_{H_a|S}}^{[h_a=i]} \quad (2)$$

$$P(H_b|S) = Cat(K, \boldsymbol{p_{H_b|S}}) = \prod_{i_{H_b|S}=1}^{K} p_{i_{H_b|S}}^{[h_b=i_{H_b|S}]} \quad (3)$$

H_a, H_b가 주어졌을 때 같은 근원일 확률은 P $P(C|H_a, H_b)$로 표현할 수 있고 다른 근원일 확률은 $P(\neg C|H_a, H_b) = 1 - P(C|H_a, H_b)$로 표현할 수 있다. 베



[그림 1] 제안하는 모델의 전체 구조도. o_a, o_b는 각각 센서 a, b로부터의 센서 정보를 나타내고 h_a, h_b는 각각 센서 정보로부터 얻은 추상적인 정보(클래스 레이블)를 나타낸다. S, S_a, S_b는 각각 같은 근원일 경우의 실제 근원 정보(클래스 레이블), 다른 근원일 경우의 각각의 근원 정보(클래스 레이블)를 나타내고 C는 같은 근원(Common), 다른 근원(¬ Common)에 대한 정보를 나타낸다.

이즈 법칙을 적용하면

$$P(C|H_a,H_b) = \frac{P(H_a,H_b|C)P(C)}{P(H_a,H_b)} \tag{4}$$

로 계산할 수 있다. $P(H_a,H_b|C)$는 같은 근원일 경우의 결합 확률이고 $P(C)$는 H_a, H_b가 같은 근원을 가질 사전확률(Prior probability)이다. $P(H_a,H_b)$는 두 변수에 대한 결합 확률인데 근원이라 정의한 C에 대해 식 (5)와 같이 Marginalize 함으로써 구할 수 있다.

$$\begin{aligned} P(H_a,H_b) = {} & P(C)P(H_a,H_b|C) \\ & + P(\neg C)P(H_a,H_b|\neg C) \end{aligned} \tag{5}$$

여기에서 $P(H_a,H_b|C)$와 $P(H_a,H_b|\neg C)$는 다음과 같이 구할 수 있다.

$$\begin{aligned} P(H_a,H_b|C) &= \sum_{i=1}^{K} P(H_a,H_b|s=i)P(s=i) \\ &= \sum_{i=1}^{K} P(H_a|s=i)P(H_b|s=i)P(s=i) \end{aligned} \tag{6}$$

$$\begin{aligned} p(H_a,H_b|\neg C) &= \sum_{i=1}^{K} P(H_a|s_a=i)P(s_a=i) \\ &\times (\sum_{j=1}^{K} P(H_b|s_b=j)P(s_b=j)(1-[s_a=s_b])) \\ &, where\ [X]\begin{cases}1 \text{ if } X \text{ is true}\\ 0 \text{ otherwise}\end{cases} \end{aligned} \tag{7}$$

위의 식 (5), (6), (7)은 이산 분포이기 때문에 조합함으로써 H_a, H_b가 주어졌을 때 같은 근원일 확률 $P(C|H_a,H_b)$를 구할 수 있다.

$P(C|H_a,H_b) \geq 0.5$ 일 경우, 주어진 데이터에 대한 클래스 레이블 추정값 $\hat{S}$는 식 (8)과 같다.

$$\begin{aligned} \hat{S} &= \underset{i}{\arg max}\, P(s=i|H_a,H_b) \\ &= \underset{i}{\arg max}\, P(H_a,H_b|s=i)P(s=i) \\ &= \underset{i}{\arg max}\, P(H_a|s=i)P(H_b|s=i)P(s=i) \end{aligned} \tag{8}$$

$P(C|H_a,H_v) < 0.5$ 일 경우엔

$$\begin{aligned} \hat{S}_a &= \underset{i}{\arg max}\, P(s=i|H_a) \\ \hat{S}_b &= \underset{i}{\arg max}\, P(s=i|H_b) \end{aligned} \tag{9}$$

중 높은 확률 값을 갖는 $\hat{S}$로 추정하게 된다. $P(S|H_a), P(S|H_b)$는 베이즈 법칙을 이용하여 각각

$$\begin{aligned} P(S|H_a) &= \frac{P(H_a|S)P(S)}{P(H_a)} \\ P(S|H_b) &= \frac{P(H_b|S)P(S)}{P(H_b)} \end{aligned} \tag{10}$$

를 통해 구할 수 있다.

3. 데이터

3.1 데이터 수집

본 논문에서는 2명의 실험 참가자가 각 14일 동안 구글 글래스를 착용하고 일상 생활을 하면서 수집한 비디오 및 오디오 데이터를 사용하였다. 또한 같은 기간 동안 스마트폰을 이용하여 사용자의 활동(activity) 내역을 기록, 클래스 레이블로 활용하였다. 수집한 데이터 중 일반적인 분류 상황을 고려하고 물리적으로 걸리는 시간을 줄이기 위해 총 데이터 중 일부인 5개의 클래스 레이블(Meal, Bus, Car, Research, Rest)에 대해 60000개(12000개/클래스 레이블, 1개/s)의 데이터를 임의로 추출해 training set 50000개(10000개/클래스 레이블), test set 10000개(2000개/클래스 레이블)로 실험을 진행하였다.

3.2 데이터 전처리

비디오 데이터의 경우 구글 글래스로부터 녹화된 1인칭 시점의 비디오로부터 초당 한 장씩의 이미지를 추출한 후, ImageNet Database를 학습한 AlexNet (Deep

CNN)으로 특징 벡터를 추출하였다.[5]. 최근 연구 결과에 의해 CNN을 이용하여 추출된 특징 벡터를 이용하여 전이학습이 가능하다고 밝혀졌다.[6] 본 실험에서는 시각 정보로부터 피험자의 활동을 구분할 때 시각 정보 내의 물체 정보가 중요할 것이라는 가정 하에 기존 전이학습 방법과는 달리 ImageNet의 마지막 층인 클래스에 대한 확률값 집합을 특징 벡터로 사용하였다. 또한 계산 복잡도를 줄이기 위해 주성분분석(PCA)를 이용하여 64차원으로 축소하였다.

오디오 데이터의 경우, 사람의 소리 인지 주파수를 반영하는 것으로 알려져 있는 MFCC 계수(Mel-frequency cepstral coefficient)를 이용하여 초당 13차원×15개의 특징 벡터를 추출하였다. [표 1]에 사용한 데이터 정보를 정리하였다.

3.3 데이터 특징 분석

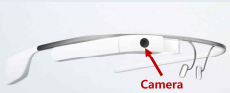
본 실험에서 사용한 데이터는 제약조건이 전혀 없는 실제 생활 환경에서 수집한 데이터이다. 따라서 기존 많은 연구에서 수행한 벤치마크 데이터나 제약이 있는 환경 내에서 실험을 통해 수집한 데이터에 비해 훨씬 자유도가 높고 잡음이 많다.

또한 각 센서에서 추출된 특징 벡터는 추상화 정도의 차이가 있다. 비디오로부터 추출된 이미지 데이터의 경우, CNN으로 학습된 특징 벡터를 사용하였을 때 가장 강력한 성능을 나타낸다. 이는 해당 특징 벡터의 추상화 정도가 높음을 의미한다. 하지만 오디오 특징 벡터로 사용한 MFCC의 경우 CNN을 이용한 것에 비해 상대적으로 추상화 정도가 낮기 때문에 본 실험에서는 비디오 데이터가 지배적인 성향을 띈다. 실제로 비디오 특징 벡터만 이용하여 Naive Bayes 분류기를 학습시켰을 경우에 정확도는 43.03%, 오디오 특징 벡터만 이용하여 Naive Bayes 분류기를 학습시켰을 경우에는 32.23%의 정확도를 보였다.

4. 실험결과

주어진 데이터에 대해서 비디오와 오디오 데이터를 통합하여 Naive Bayes 분류기를 학습했을 경우, 그리고 두 데이터를 이용하여 제안하는 모델으로 학습했을 경우를 [표 2]에 정리하였다. 전자의 경우 비디오 정보와 오디오 정보를 동일하게 비중을 두고 합친 모델로 특정 클래스 레이블에서 주어진 데이터가 있을 때 어떠한 데이터가 더 좋은 영향을 끼치는지 고려되지 않았다고 볼 수 있다. 하지만 제안하는 모델에서는 특정 클래스 레이블마다 어떠한 데이터가 좋은 영향을 끼치는 지를 확률적으로 비중을 계산하여 결정하고, 또한 합쳐야 되는지에 대해서도 확률적으로 계산하여 결정하기 때문에 앞의 경우

[표 1] 사용된 데이터 정보

	비디오 데이터	오디오 데이터
수집 장비	 구글 글래스 카메라	 구글 글래스 마이크
데이터 특징 벡터	64차원 Deep Convolution Neural Network 특징 벡터	195(13× 15)차원 MFCC 특징 벡터
클래스 레이블	사람의 활동(activity) {식사, 버스, 승용차, 연구, 휴식}	사람의 활동(activity) {식사, 버스, 승용차, 연구, 휴식}
시간 해상도	1개/s	1개/s

[표 2] 실험 결과. 비디오+오디오를 Naive Bayes 분류기로 학습했을 경우의 정확도와 비디오+오디오를 제안하는 모델로 학습했을 경우의 정확도를 나타낸다.

	비디오+오디오 Naive Bayes	비디오+오디오 제안하는 모델
정확도 (%)	34.26 %	39.89 %

보다 더 좋은 성능을 얻을 수 있었다.

비디오만 이용하여 Naive Bayes 분류기를 학습했을 경우, 정확도가 가장 높았는데 이러한 이유는 앞 절에서 설명하였듯이 비디오 데이터를 이용해 추출한 특징 벡터가 지배적인 성향을 띄기 때문으로 해석된다.

5. 논의 및 결론

본 논문에서는 사람의 다중 신호 처리 방식을 모방하여 웨어러블 센서로부터 입력되는 다중 센서 데이터를 효율적으로 통합하여 처리할 수 있는 모델을 제안하였다. 해당 모델은 다중 센서 정보 중 같은 근원으로부터 발생된 정보는 통합하여 처리하고 그렇지 않은 경우엔 분리하여 처리한다. 해당 모델을 실제 생활 환경에서 구글 글래스로 수집한 1인칭 시점의 비디오, 오디오 데이터를 이용해 실험하였다. 실험 결과 동등하게 통합하는 결과에 비해 높은 성능을 보였다.

제안하는 모델은 센서 통합을 위한 초기 모델로서 한계점이 분명하다. 계산의 간편함을 위해 2개의 센서에 대해서만 설명하였고 각 센서 처리 모듈로서 분류기를 사용하였다. 또한 센서 데이터는 시계열 데이터이므로 시간에 대한 정보를 고려해야 하지만 여기서는 시간에 대한 정보를 무시하였다. 향후 연구로는 시계열 정보를 고려한 추상적인 정보를 학습할 수 있는 센서 처리 모듈을 고안하고 3개 이상의 센서에 대해 통합하는 방법을 연구할 것이다.

감사의 글

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)을 받아 수행된 연구임.

참고문헌

[1] B.-T. Zhang, Ontogenesis of agency in machines: A multidisciplinary review, *AAAI 2014 Fall Symposium on The Nature of Humans and Machines: A Multidisciplinary Discourse*, Arlington. 2014.

[2] Bruno, Barbara, et al, Analysis of human behavior recognition algorithms based on acceleration data, *Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE*, 2013.

[3] Jacob, Robert JK, and Keith S. Karn, Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind* 2.3 (2003): 4.

[4] Trommershauser, Julia, Konrad Kording, and Michael S. Landy, eds. Sensory cue integration. Oxford University Press, 2011.

[5] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*. 2012.

[6] Yosinski, Jason, et al, How transferable are features in deep neural networks?, *Advances in Neural Information Processing Systems*, 2014.