음향-영상-위치 통합정보의 앙상블학습을 이용한 장소인식*

이충연[1○], 이범진[1], 온경운[1], 하정우[1], 강우성[2], 장병탁[1]
[1]서울대학교 컴퓨터공학부, [2]삼성전자 DMC 연구소
{cylee, bjlee, kwon, jwha}@bi.snu.ac.kr, ws.kang@samsung.com, btzhang@bi.snu.ac.kr

Place Recognition using Ensemble Learning of Audio-Vision-Location Integrated Information

Chung-Yeon Lee[1○], Beom-Jin Lee[1], Kyoung-Woon On[1], Jung-Woo Ha[1], Woo-Sung Kang[2], Byoung-Tak Zhang[1]
[1]School of Computer Sci. and Eng., Seoul National University, [2]DMC R&D Center, Samsung Electronics

요 약

본 논문에서는 음향-영상-위치 통합 정보 기반의 장소 인식 방법을 제안한다. 이 방법은 음영 지역에서의 정확도 감소나 추가 하드웨어 필요 등 기존 위치 정보 인식 방법이 가지는 제약을 극복 가능하고, 지도상의 단순 좌표 인식이 아닌 논리적 위치 정보 인식을 수행 가능하다는 점에서 해당 위치와 관련된 특정 정보를 활용하여 다양한 생활편의를 제공하는 위치 기반 서비스를 수행하는데 보다 효과적인 방법이 될 수 있다. 제안하는 방법에서는 음향, 영상, 위치 정보들과 같이 서로 다른 특성을 가진 이종 센서 데이터들로부터 추출된 특징 벡터들을 학습하기 위해, 각 데이터로부터 추출한 특징 벡터들을 각각 다른 분류기를 통해 학습한 후, 그 결과를 기반으로 최종적인 하나의 분류 결과를 얻어내는 앙상블 기법을 사용한다. 실험 결과에서는 각각의 데이터를 따로 학습하여 분류한 결과와 비교하여 높은 성능을 보였다. 또한 사용자 상황인지 기반 서비스의 성능 향상을 위한 방법으로서 제안하는 모델의 스마트폰 앱 구현을 통한 활용 가능성에 대해 논의하였다.

1. 서 론

최근 스마트 모바일 기기 사용의 대중화와 위치 기반 서비스(Location-based Service, LBS)의 활성화로 스마트 모바일 장비를 기반으로 하는 위치 정보 서비스의 중요성이 증대되고 있다. 그러나 위치 정보 인식을 위한 기존의 방법들은 대부분 경제성이나 편의성, 정확도의 부족으로 실용화가 쉽지 않은 상황이다.

위치 정보 인식을 위한 기존 방법들은 인식 가능 범위에 따라 GPS 및 이동통신망 기반의 매크로 위치 인식, 적외선이나 초음파 등을 이용하는 마이크로 위치 인식, 그리고 이동 레퍼런스 노드들 간의 연결성으로 위치를 계산해내는 Ad-Hoc 위치 인식 방법으로 분류된다[1-3]. 그러나 이 방법들은 신호가 감쇄되는 음영 지역에서 정확도가 감소하고, 추가적인 하드웨어를 필요로 하는 등의 제약이 존재한다. 또한 측정된 좌표를 실제 서비스에서 이용하기 위해 해당 위치의 논리적 정보를 개별적으로 맵핑해야 하는 어려움이 있다. 한편, 지능형 모바일 로봇의 자율이동기능 구현을 위해 영상 정보를 분석하여 위치를 파악하는 기술인 컴퓨터 비전(Computer Vision) 기반 위치 인식 기법도 활발히 연구되고 있다[4]. 이 방법은 로봇이 사람처럼 집안의 가구 위치와 벽면 각도, 바닥이나 천정, 조명의 특징 등을 인식해서 스스로 이동 경로를 설정하는 가장 고차원적인 기술이다.

* 본 논문은 2014년도 삼성전자의 지원을 받아 수행된 연구이며, 정부(미래창조과학부 및 산업통상자원부)의 재원으로 한국연구재단(NRF-2010-0017734-Videome)과 한국산업기술평가관리원(KEIT-10035348-mLife)의 지원을 일부 받았음.

본 논문에서는 좌표 정보 외에 해당 위치에서의 시각 및 청각 신호 등 멀티모달(Multi-modal)의 환경 정보들을 이용하는 사람의 위치 인식 메커니즘을 모방한 음향-영상-위치 통합 정보 기반의 위치 정보 인식 방법을 제안한다. 이 방법은 앞서 기술한 기존 위치 정보 인식 방법과는 달리, 지도상의 단순 좌표 인식이 아닌 논리적 위치 정보, 즉 장소(Place) 인식을 수행 가능하다는 점에서 해당 위치와 관련된 특정 정보를 활용하여 위치 기반 서비스의 성능을 향상하는 데 효과적인 방법이 될 수있다.

음향, 영상, 위치를 포함하는 서로 다른 특성을 가진 이종(Heterogeneous) 데이터들을 함께 이용하기 위해, 제안하는 방법에서는 각 데이터로부터 추출한 특징 벡터들을 서로 다른 분류기를 통해 학습하고 그 결과를 가중치 기반의 voting method를 통해 최종적인 하나의 분류 결과를 얻어내는 앙상블 학습(Ensemble Learning) 방법을 사용한다[5]. 앙상블 학습은 주어진 학습 데이터를 가장 잘 설명하는 가설 하나를 찾는 대신 가설들의 집합을 만들고 이 집합으로 투표를 하여 예측하는 방법으로, 유전자 정보 분석[6], 기후 변화 예측[7], 영화 추천 시스템[8] 등에서 이미 이종 데이터 학습에 사용된바 있다.

MFCC (Mel-frequency Cepstral Coefficients) [9]와 SIFT (Scale-Invariant Feature Transform) 기술자 [10]가 각각 영상과 음향의 특징 벡터를 추출하는데 사용되며, 위치 데이터는 장소별 확률을 계산을 위한 가중치로 사용된다. 제안하는 방법을 검증하기 위해 실내 장소(사무실, 강의실, 식당)와 실외 장소(버스, 지하철, 자가용)에서 각각 데이터를 획득하고, 장소 인식 성능 평가를 수행하였다. 그림 1은 제안하는 앙상블 기반 장소인식 기법의 전체 구조를 도식화한 것이다.

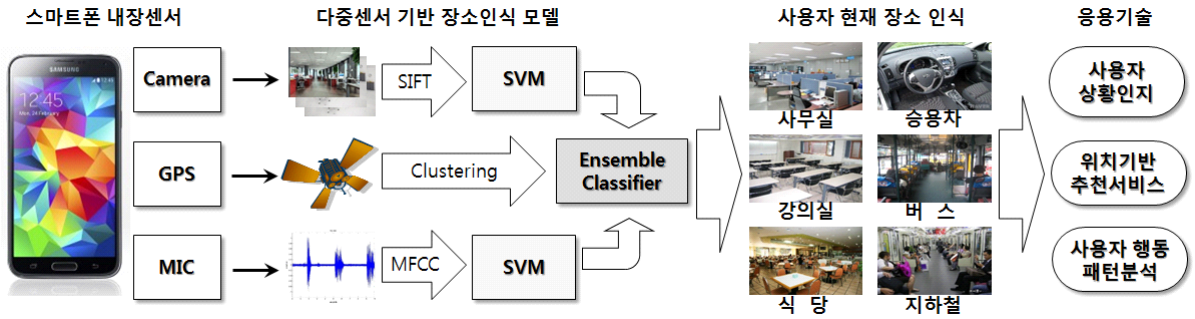

그림 1. 스마트폰에 내장된 다중 센서 기반 사용자 장소인식 기법 전체 흐름도

2. 연구 내용 및 방법

2.1 음향 데이터 전처리 및 특징 추출

음향 데이터로부터 특징 벡터를 추출하기 위해 사람의 소리 인지 주파수를 반영하는 MFCC 계수를 이용한다. 먼저 음향 데이터를 44100 Hz의 샘플링 주파수로 1초 동안 녹음한 후, 2048개 샘플로 구성된 프레임을 1024개 간격으로 이동시키면서 이전 프레임과 50% 중첩되는 방식으로 분할한다. 이렇게 총 42개 프레임으로 나뉜 데이터는 주파수 영역으로 변환할 때 발생하는 프레임별 양 끝단의 불연속 지점에서의 오류를 최소화하기 위해 각 프레임에 식 (1)과 같은 해밍 윈도우(Hamming Window)를 적용한다. 여기서 N은 프레임의 길이이다.

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi}{N}n) \quad (1)$$

각 프레임의 데이터를 FFT (Fast Fourier Transform)를 이용하여 주파수 영역으로 변환한 후, Mel-scale의 필터 뱅크(Filter Bank)에 통과시켜 파워스펙트럼을 구한다. 입력 주파수를 f라고 할 때, Mel-scale의 주파수 $M(f)$를 다음 식 (2)와 같이 구할 수 있다.

$$M(f) = 2595\log_{10}(1 + \frac{f}{700}) \quad (2)$$

Mel-scale 필터 뱅크를 통과한 데이터에 식 (3)과 같이 로그 함수와 DCT (Discrete Cosine Transform)를 수행하여 12개의 주파수 특성과 하나의 프레임 로그 에너지 특성으로 구성된 총 13차의 MFCC 계수 $A(n)$을 추출한다. 이때 $Q_k(k = 1,2,...,K)$는 필터 뱅크의 출력 데이터, N은 프레임 내 샘플 개수이다.

$$A(n) = \sqrt{\frac{2}{k}}\sum_{k=1}^{K}\log(Q_k)\cos\left(n(k-0.5)\frac{\pi}{k}\right) \quad (3)$$

최종적으로 음향 데이터로부터 1초마다 42개의 특징 벡터들이 구해지며, 모두 학습 데이터 집합으로 사용된다.

2.2 영상 데이터 전처리 및 특징 추출

영상 데이터는 480×640 pixels (72 dpi)의 해상도를 가지는 이미지를 임의의 각도로 촬영하여 획득하며, 이때 조리개 개방 및 노출 정도는 카메라가 자동으로 조정한다. 이후 영상 데이터의 크기를 120×160 pixels로 축소하고 그레이스케일(Grayscale) 영상으로 변환한 후, SIFT 기술자를 사용하여 특징 벡터를 추출한다.

추출된 전체 SIFT 기술자들은 k-means 클러스터링을 적용하여 200개 크기의 Bag-of-Words를 특징 벡터로 표현되고 이에 따라 각 이미지는 200개 시각단어의 히스토그램으로 정의된다.

2.3 위치 데이터 전처리 및 특징 추출

위치 데이터는 스마트폰에 탑재된 A-GPS 센서를 이용하여 영상 데이터와 함께 지오태그(GeoTag)의 형태로 기록된다. 영상 데이터의 EXIF (Exchangeable Image File Format) 메타 데이터에 기록되어 있는 지오태그로부터 도분초(*dd°mm'ss"*) 단위의 GPS 좌표를 추출하고, 이를 도(degree) 단위(*DD.DDDD*)의 데이터로 변환한다. 다음으로 전체 GPS 데이터를 위도와 경도가 각각 0.001로 그리드 영역들로 구분지어 클러스터링한 후, 수식 (4)와 같이 각 영역들에 포함되어 있는 장소 레이블 개수를 카운트하고, 해당 영역들에 각 장소에 대한 확률 가중치를 계산하여 부여한다.

$$G(c,p) = \left\{\sum_{i\in N}\delta(p, c_p(i))\right\} \;/\; N_c \quad (4)$$

식 (4)에서 c는 장소 클러스터를, p는 장소 레이블을 나타내며, N은 전체 GPS 데이터 수, N_c는 클러스터 c에 포함되어 있는 GPS 데이터 수를 나타낸다.

2.4 장소 인식 앙상블 학습 모델

음향, 영상, 그리고 위치 데이터는 각기 다른 형태의 특징 벡터와 모델을 가지고 있기 때문에, 단일 분류기를 통해 학습하는 방법이 아닌 데이터별로 별개로 학습된 분류기의 예측값들을 결합하는 앙상블 방법을 사용한다.

먼저 음향과 영상 데이터의 특징 벡터들을 각각 SVM 분류기를 이용하여 분류한다. 단, 음향 데이터는 42개의 서로 다른 분류 결과들 중 추정확률(probability estimate) 합이 가장 큰 장소 레이블을 선택하고, 해당 레이블이 차지하는 비율을 가중치로 사용한다. 영상 데이터로부터 분류된 장소 레이블은 추정확률값을 가중치로 사용한다.

최종적으로 식 (5)와 같이 위치 데이터의 영역별 장소 확률 가중치를 음향과 영상 데이터 분류 결과의 가중치에 곱한 후, 두 가중치를 비교하여 높은 가중치를 가지는 장소 레이블 p_{m*}을 구한다.

$$m* = \arg max_m\{\alpha G(c, p_m) w_m\},\ m \in \{audio, vision\} \quad (5)$$

식 (5)에서 p_m은 음향과 영상 데이터에서 각각 분류된 장소 레이블을, w_m는 분류 결과의 가중치를 나타낸다.

3. 실험 결과 및 논의

본 실험에 사용된 데이터는 6명의 학생 및 연구원들이 각자 스마트폰을 이용하여 촬영한 사진 3000장, GPS 좌표 9000개, 그리고 3000초의 녹음된 오디오 파일들을 사용하였다. 분류를 위한 장소는 사용자가 주로 일상을 보내는 실내외 장소들을 고려하여 사무실, 강의실, 식당, 버스, 지하철, 자가용으로 총 6개 장소를 선택하였다.

실험에서는 먼저 전체 데이터를 훈련 데이터와 검정 데이터의 두 그룹으로 나누며, 이때 검정 데이터는 전체 데이터에서 임의로 추출된 20%의 데이터를 사용하고, 나머지를 훈련 데이터로 사용하였다. 하나의 데이터 샘플은 각각 사진 1장으로부터 추출된 영상 데이터 특징 벡터(k=200), 3개의 GPS 좌표값, 그리고 1초 길이의 환경 녹음 사운드로부터 추출된 42개의 MFCCs로 구성되었다.

제안한 방법을 사용하여 분류한 결과는 그림 2와 같다. 분류 성능의 비교를 위해 각 데이터를 따로 분류한 결과와 영상-음향 데이터 분류기 2개를 사용하여 분류한 결과를 함께 확인하였다. 이때 음향과 영상 데이터는 각 특징 벡터를 SVM을 이용하여 분류하였으며, 위치 데이터는 위에서 기술한 방법으로 생성된 장소 클러스터와 검정 데이터를 비교하여 분류하였다. 영상-음향 데이터 분류는 샘플별 분류 결과에서 추정확률값이 높은 결과를 선택하는 방식으로 분류하였다.

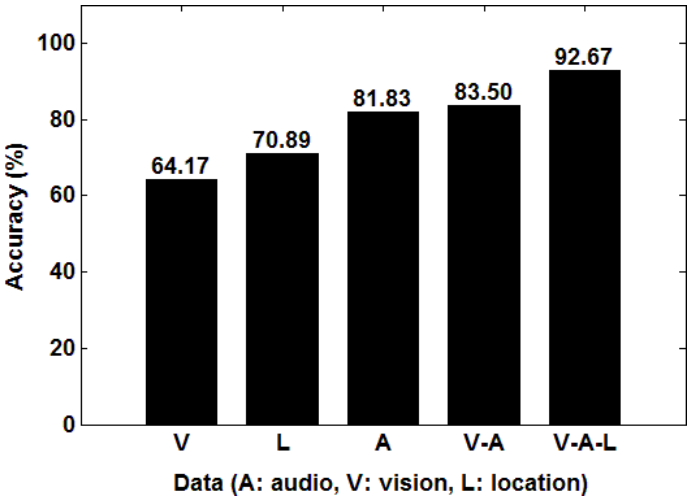

그림 2. 학습 데이터별 장소 인식 결과 비교

실험 결과에서는 영상 데이터와 음향 데이터를 함께 사용한 것이 단일 데이터 분류기들보다 높은 인식률을 보였으며, 세 가지 데이터를 함께 사용하는 경우 월등히 높은 인식률을 나타냄을 확인할 수 있었다.

이때 영상 데이터가 위치/음향 데이터에 비해 상대적으로 낮은 인식률을 나타냄은 각 장소에서 촬영된 사진들이 조도 변화나 흔들림 등 외부 요인에 의한 잡음(Image Variability)을 포함하였기 때문으로 볼 수 있으며, 위치 데이터는 동일 건물 내에 위치한 사무실, 강의실, 식당을 잘 분류하지 못하였음이 확인되었다. 한편, 음향 데이터의 경우 서로 다른 장소(예: 버스 & 자동차)에서 녹음된 데이터 내에 유사한 음향 신호가 포함되어 있기 때문인 것으로 추측해볼 수 있다(Perceptual Aliasing) [11].

따라서 제안한 방법은 단일 데이터 분류시 문제가 되는 요인들을 서로 다른 데이터를 활용하여 상보적으로 완화시킴으로써 인식률을 향상시킨 것으로 해석된다.

4. 결론 및 향후 연구

본 논문에서는 스마트폰에 장착된 카메라, 마이크 GPS 센서를 통해 입력받은 멀티모달 데이터를 이용하여 장소 인식을 수행하는 방법을 제안하였다. 실험 결과는 제안하는 방법이 단일 데이터를 사용한 것보다 월등히 높은 인식률을 보였다. 이밖에도 제안하는 방법을 통해 기존 GPS 기반의 방법이 불가능한 실내 장소 인식(Indoor Place Recognition)이 가능하다. 향후 지도 정보 서비스와의 연계 및 스마트폰 앱 구현을 통해 사용자의 실시간 상황인지 및 위치 기반 서비스의 성능 향상을 위한 효과적인 방법이 될 것으로 기대된다.

참 고 문 헌

[1] K. Whitehouse and D. Culler, "Macro-calibration in sensor/actuator networks," *Mobile Networks and Applications*, Vol. 8, No. 4, pp. 463-472, 2003.

[2] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, Vol. 34, No. 8, pp. 57-66, 2001.

[3] H. Koyuncu and S. H. Yang, "A survey of indoor positioning and object locating systems," *International Journal of Computer Science and Network Security*, Vol. 10, No. 5, pp. 121-128, 2010.

[4] D. C. Herath, S. Kodagoda and G. Dissanayake, "A Two-tier Map Representation for Compact stereo-vision-based SLAM," *Robotica*, Vol. 30, No. 2, pp. 245-256, 2012.

[5] T. Dietterich, "Ensemble methods in machine learning," In *Multiple Classifier Systems*, pp. 1-15, Springer Berlin Heidelberg, 2011.

[6] D. Marbach et al., "Wisdom of crowds for robus gene network inference," *Nature Methods*, Vol. 9, No. 8, pp. 796-804, 2012.

[7] P. L. Vidale, D. Lüthi, R. Wegmann, C. Schär, "European summer climate variability in a heterogeneous multi-model ensemble," *Climatic Change*, Vol. 81, No. 1, pp. 209-232, 2007.

[8] X. Shi, J-F. Paiement, D. Gragier and P. S. Yu, "Learning from heterogeneous sources via gradient boosting consensus," Proc. of the 12th SIAM Int`l Conf. on Data Mining (SDM 2012), pp. 22-235, 2012.

[9] M. Xu et al., "HMM-based audio keyword generation," In *Advances in Multimedia Information Processing*. Springer, 2004.

[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, 2004.

[11] B. Kuipers & P. Beeson, "Bootstrap learning for place recognition," Proc. of 18th National Conference on Artificial Intelligence (AAAI 2002), pp. 174-180, 2002.