실시간 장소 인식 시스템의 설계 및 구현

오수진○ 남양희
이화여자대학교 디지털미디어학부
bluerhino@hanmail.net, yanghee@ewha.ac.kr

Real-Time Place Recognition System Design and Implementation

Sujin Oh○ Yanghee Nam
Division of Digital Media, Ewha Womans University

모바일기기들을 중심으로 컴퓨팅능력이 통합 확장됨으로써 개인화된 서비스의 필요성이 커지고 있으며, 특히 사용자의 위치와 상황에 따라 적합한 정보를 제공하기 위한 방법으로 모바일 혼합/증강현실에 대한 관심과 활용도 또한 증가되고 있다. 사용자에게 필요한 정보를 적절하게 서비스하기 위해서는 사용자가 속해있는 공간에 대한 정보 및 사용자의 위치와 방향을 파악함으로써 상황을 인지하는 기술이 필연적으로 요구된다. 그동안 장소인식 기술에 관련한 많은 연구가 있었으나 모바일 혼합현실에 적용된 사례는 드물었다. 실내 환경에 적합하며 보다 정밀한 위치인식이 요구되는 시스템에서는 센서기반 방식보다 카메라기반 방식이 주로 활용되는데, 영상신호만을 이용하여 장소를 인식하는 것은 카메라의 움직임에 따른 블러링, 영상잡음, 조명상태에 민감한 영향을 받게 된다. 이를 극복하기 위하여 그동안 로봇비전 및 웨어러블 컴퓨팅분야에서 다양한 연구가 이루어져왔다.

Torralba는 한대의 카메라를 이용한 장소인식 시스템을 제안하였는데, 영상으로부터 steerable pyramid방법을 적용하여 전역적인 특징을 추출하고 HMM 모델을 구축하여 사용자가 현재 각각의 장소에 위치할 확률을 구하였다. 그러나, 이러한 전역특징 추출방법은 장면의 부분적인 폐색이나 잡음에 민감하여 올바르게 장소를 인식하는 것이 어렵다[1]. Li는 영상으로부터 지역적인 특징을 추출하기 위한 방법으로 scale-invariant keypoints(SIFT) 방식을 이용하였으며, 유효한 특징데이터의 양을 줄임으로써 인식속도를 향상시키고 좀더 넓은 인식공간에 대처할 수 있는 능력을 한 단계 높여주었다. 그러나, 물체인식에 주로 사용되는 SIFT 방식은 계산량이 많아 실시간성이 요구되는 모바일 증강현실에 적용하기에는 한계가 있다[2]. Min은 여러 대의 카메라를 동시에 사용하여 네 방향의 영상을 얻어냄으로써 인식율을 높이는 방법을 제안하였으나 이는 단일카메라를 사용하는 모바일 시스템에는 적합하지 않다[3]. 그 밖에 장소인식에 미리 학습된 물체들의 문맥정보를 활용하여 인식율을 높이는 연구들이 이루어졌으나 영상으로부터 물체를 인식하기 위해 많은 시간이 소요되어 실시간으로 장소를 인식하는 시스템에 적용하는데 한계가 있다[4].

한정된 컴퓨팅 리소스로 구성된 모바일기기에서 영상신호로부터 사용자의 위치와 방향을 알아낸 후 실시간으로 서비스될 정보를 디스플레이하기 위해서는 계산량이 많은 특징추출방법을 이용했던 기존의 연구들과 달리 단일카메라를 이용하면서 실시간성을 높이는 알고리즘이 필요하다. 본 논문에서는 사용자의 상황문맥에 맞는 증강정보를 가시화하여 모바일 AR Annotation & Guide System에 활용될 수 있는 실시간 비디오기반 장소인식 시스템을 제안하고자 한다. 모바일 증강현실에 적용 가능하도록 안정된 인식률과 실시간성을 높이기 위해 본 논문에서는 실시간 비디오 검색의 분야에서 활용되었던 방식인 color 와 texture 정보를 이용한 영상 특징추출 방법[5]을 사용하였다. 본 논문의 장소 인식 시스템은 실시간으로 카메라를 통해 들어오는 영상신호에서 특징을 추출한 후, 장면 분류에 따른 관측확률을 추정하고 은닉 마르코프 모델(HMM)을 이용해 현재 사용자가 위치할 확률이 가장 높은 장소를 결과로 알려주는 시스템이다. 모델에 입력값으로 사용되는 장소이동확률은 장소간의 위치관계를 이용하여 보다 정확하게 구할 수 있다. 또한, 오프라인에서 수집된 레이블링된 장면이미지가 학습데이터로 사용된다. 장소인식결과에 따라 모바일 환경에서는 사용자 문맥에 맞는 정보를 디스플레이할 수 있게 된다. 그림 1은 시스템의 전체 구조도이다.

장소인식을 위한 그래피컬 모델중의 하나로 HMM을 이용하는데, 이 모델에서 은닉된 상태는 인식하고자 하는 사용자의 장소, 관측값은 카메라로부터의 영상신호가 된다. 시간 t에서 사용자가 위치할 장소를 Q_t라 하고, 전체 이미지 특징벡터를 Z_t라 할 때, 현재까지의 관찰된 영상정보들에 대하여 현재 사용자가 각각의 장소에 위치할 확률은

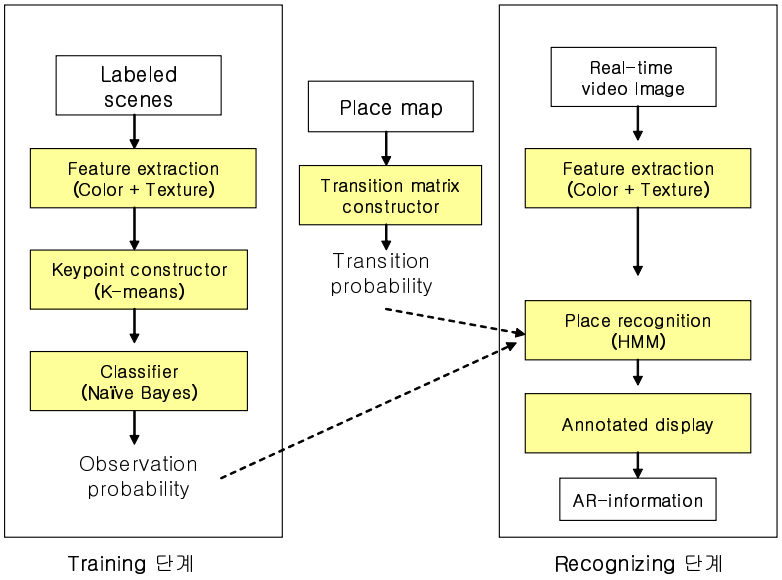

그림1. 전체 구조도

$P(Q=q|Z_{1:t})$이라 하며, 관측확률과 장소이동확률을 이용하여 재귀적으로 계산할 수 있다[1]. 영상으로부터 특징을 추출하는 것은 다음과 같은 방법을 사용한다. 먼저, 장면 이미지를 10*10 grid로 나누어 patch로 구분한다. 각 patch를 묘사하는 특징벡터는 두 종류의 값으로 구성되는데 하나는 HSI 칼라 히스토그램, 또 하나는 texture 특징값이다. Patch의 color정보는 16 bins의 hue 히스토그램, 6 bins의 intensity 히스토그램으로 표현한다. texture 정보는 patch에 N*N 마스크를 이용하여 intensity의 공분산행렬을 구한 후, 공분산과 고유값으로부터 3개의 특징값을 계산하여 이용한다. 즉, 한 patch당 color와 texture정보를 합하여 25차원의 이미지 특징벡터를 만들어 사용한다[5]. HMM에 필요한 Observation Probability(관측확률)은 표본 데이터로부터의 학습을 통해서 추정되는데 이는 레이블링된 학습 데이터를 장소에 따라 분류화시킴으로써 계산할 수 있다. 본 장소인식 시스템에서는 이를 위해 the bags of keypoints method를 이용한다[6].

본 논문의 장소 인식 시스템을 테스트하기 위하여 대학 내의 한 건물 내에서 세부장소를 인식하는 실험을 하였다. 총 12곳의 장소에서 90장의 레이블링 이미지를 훈련데이터로 이용하였으며 실험에 필요한 영상은 320*240의 영상데이터 시퀀스를 사용하였다. 본 시스템의 프로세싱 타임은 약 15 f/s의 성능과 약 76%의 인식율을 보였다. 실험결과, 적합한 인식률을 유지하면서 비교적 간단한 특징추출계산을 통해 실시간성이 높아짐을 확인할 수 있었다. 향후 연구에서는 외부조명상태에 강건한 알고리즘의 개발이 추가되어야 할 필요가 있고, 모바일 플랫폼에 적용 가능한 라이브러리의 개발과 장소인식 시스템을 활용할 수 있는 모바일 증강현실 응용 애플리케이션의 개발이 뒤따라야 할 것이다. (본 연구는 서울시 산학연 협력사업의 지원에 의한 것임)

참고 문헌

[1] A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition", IEEE Int. Conf. on Computer Vision, 2003

[2] Fayin Li, J.Kosecka, "Probabilistic Location Recognition using Reduced Feature Set", IEEE Int. Conf. on Robotics and Automation, 2006

[3] 민경민, 이성훈, 김동호, 김진형, "Nonstaionary HMM을 이용한 다중 카메라 기반 장소 인식", HCI 2007

[4] S. Kim, I. Kweon, "Collaborative Place and Object Recognition in Video using Bidirectional Context Information", 제1회 한국지능로봇 하계종합 학술대회 논문집 2006

[5] M. Israel, E. L. van den Broek, P. Van der Putten, and M. J. Den Uyl. "Automating the Construction of Scene Classifiers for Content-Based Video Retrieval", Int. Workshop on Multimedia DataMining, 2004

[6] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, "Visual Categorization with Bags of Keypoints", In European Conference on Computer Vision, 2004