이동형 정보 증강 시스템을 위한 실시간 장소 인식

(Real-Time Place Recognition for Augmented Mobile Information Systems)

오 수 진 † 남 양 희 ††

(Sujin Oh) (Yanghee Nam)

요 약 이동 중 사용자에게 필요한 정보를 제공하기 위해서는 장소를 인지하는 기술이 필요하다. 본 논문에서는 건물 내에서 이동하면서 카메라에 의해 포착된 영상 정보를 분석하여 현재 장소를 파악하고 카메라 영상에 관련 정보를 증강하는 비디오 기반 실시간 장소인식 시스템을 제안한다. 영상의 전역적 특징을 이용한 기존 연구들은 장면의 부분적인 폐색이나 잡음에 민감하고, 물체인식을 행하는 지역적 특징 의존 방식은 계산량이 많아 실시간 적용이 어렵다. 또한, 그러한 특징들로부터 장소인식 결과를 도출하기 위해서는 통계적 그래프 기반 모델이나 베이시안 네트웍 등이 이용되어 왔는데, 전자의 경우 장소 이동의 확률을 얻기 위한 많은 통계 데이타가 필요하며, 후자는 장소 이동 문맥을 활용하지 못하므로 물체 인식 결과에만 의존하는 단점이 있다. 본 논문에서는 장소 문맥 정보를 활용하면서 영상의 지역적, 전역적 특징추출법의 결합을 통해 부분 폐색 및 잡음에 대한 전역적 방법의 민감성을 보완하고, 지역적 방법의 느린 처리속도를 보완한 시스템을 제안한다. 제안된 방법을 건물 내부를 이동하면서 장소에 대한 정보를 얻는 정보증강 시스템에 적용하여 실시간 성능을 확인하였다.

키워드 : 장소 인식, 이동형 정보 증강 시스템, 장면 분류, 베이시안 분류기, 백 오프 키포인트

Abstract Place recognition is necessary for a mobile user to be provided with place-dependent information. This paper proposes real-time video based place recognition system that identifies users' current place while moving in the building. As for the feature extraction of a scene, there have been existing methods based on global feature analysis that has drawback of sensitiveness for the case of partial occlusion and noises. There have also been local feature based methods that usually attempted object recognition which seemed hard to be applied in real-time system because of high computational cost. On the other hand, researches using statistical methods such as HMM(hidden Markov models) or bayesian networks have been used to derive place recognition result from the feature data. The former is, however, not practical because it requires huge amounts of efforts to gather the training data while the latter usually depends on object recognition only. This paper proposes a combined approach of global and local feature analysis for feature extraction to complement both approaches' drawbacks. The proposed method is applied to a mobile information system and shows real-time performance with competitive recognition result.

Key words : place recognition, mobile system, scene classification, bayesian classifier, bag of keypoints

1. 서 론

유비쿼터스 시대에 컴퓨팅 환경은 점점 더 개인화되어가고 있으며, 사용자의 상황에 맞는 서비스를 제공하는데 필요한 기술들이 요구되고 있다. 특히, PDA나 PMP, 휴대폰 등의 이동형 기기들을 중심으로 컴퓨팅 능력이 통합, 확장됨으로써 개인화된 모바일서비스의 필요성이 커지는 요즈음 사용자의 위치와 상황에 적합한 정보를 제공하기 위한 기반기술로서의 장소 인식 기술은 중요한 연구 분야이다.

장소인식 시스템을 크게 두 가지로 구분하면 GPS나 RFID등의 센서기반 방식과 비전 기반 방식이 있는데 본 논문에서는 특수 센서 등의 설치 및 유지보수 비용과 실내 환경에서 GPS 사용이 불가함을 고려하여 비전 기반의 실시간 장소 인식 문제를 다룬다. 즉, 카메라 취득 영상을 분석하여 사용자의 소속 공간을 알아내는 것이 해결해야 할 문제이다.

본 논문에서 제안하는 방법은 계산량이 많은 특징 추출 방법을 이용했던 기존의 연구들과 달리, 단일 카메라를 사용하면서 실시간성을 높인 비디오 기반 장소인식 기술로서, 비디오 영상의 전역적, 지역적 특성을 포함하는 특징 추출 방법과 장소 이동 컨텍스트를 바탕으로 장소별 특징의 통계적 분포 특성을 계산하여 장면을 분

· 이 논문은 서울시의 산학연 협력사업의 전략산업 혁신 클러스터 육성 지원사업(10581) 수행에 따른 결과임

· 이 논문은 2007 한국컴퓨터종합학술대회에서 '이동형 정보 증강 시스템을 위한 실시간 장소인식'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 이화여자대학교 디지털미디어학부
bluerhino@hanmail.net

†† 정 회 원 : 이화여자대학교 디지털미디어학부 교수
yanghee@ewha.ac.kr

논문접수 : 2007년 10월 2일
심사완료 : 2008년 4월 5일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제14권 제5호(2008.7)

류하는 방법이다. 실험을 위하여, 모바일 실내 정보 증강 시스템을 구축하고, 그 실시간 성능 및 인식 기능을 확인하였다.

2. 관련 연구

카메라 기반 장소 인식법은 카메라의 움직임에 따른 블러링, 영상잡음, 조명상태에 따라 민감한 영향을 받게 되는데 이를 극복하기 위한 다양한 연구가 이루어져왔다.

우선 영상의 특징을 추출하는 방식에는 전역적인 것과 지역적인 특징 추출법들이 있는데, steerable pyramid나 SIFT등의 추출방법은 장소인식에 비교적 좋은 결과를 나타내나 부분 폐색에의 민감함으로 인한 인식률 저하와 계산량이 많아 실시간성이 떨어지는 단점이 있다[1-3].

최근, 장면인식 문제에서 학습데이타가 부족한 것을 도메인 지식에 대한 온톨로지 표현으로 보완하는 효율적 베이시안 네트웍 구성법이 제안된 바 있으나, 많은 레이블링 작업이 요구되고 특징적 물체의 존재 여부가 장소의 증거로 사용될 수 있어야 하며, 실시간성 보다는 적은 데이타로부터의 장면 분류에 연구의 초점이 맞춰져 있다[4]. 그 밖에 미리 학습된 물체들의 문맥정보를 활용하여 인식률을 높이는 연구들이 이루어졌으나, 영상으로부터 물체를 인식하기 위해 많은 시간이 소요되는 문제가 있으며[5,6], 여러 대의 카메라를 동시에 사용하여 높은 인식률을 얻는 방법도 제시되었다[7].

본 논문에서는 실시간 검색 시스템에 종종 사용되는 컬러와 공분산 기반의 텍스처 특징에 의한 전역적 특징 추출법을 사용하되[8], 각 영상을 격자형 분할에 의한 지역 패치(local patch)들로 나누어 이를 적용함으로써 장면의 전역적인 특징과 지역적인 특징을 동시에 고려하는 방법을 제안하였으며, 이러한 분할 패치 방법은 특징점을 중심으로 한 패치보다 더 좋은 성능을 발휘한다[9]. 추출된 특징 벡터들은 백 오프 키포인트와 베이시안 분류를 이용한 방식[10]의 장면 분류기 및 장소 연결 컨텍스트 기반의 그래프 모델을 통해 장소를 인식하게 된다.

3. 실시간 장소 인식 시스템의 설계

제안된 장소 인식 시스템은 그림 1과 같이 크게 학습단계와 실시간 인식단계로 구분된다. 학습단계에서는 레이블링된 장면 이미지들로부터 특징을 추출한 후, 장소별 특징들의 관측 확률이 계산된다. 이는 컨텍스트 그래프 모델에서 장소를 표현하는 노드의 관측 확률이 된다. 또한, 공간의 위치관계도(place-map)에 의해 장소 간 이동 확률이 설정된다.

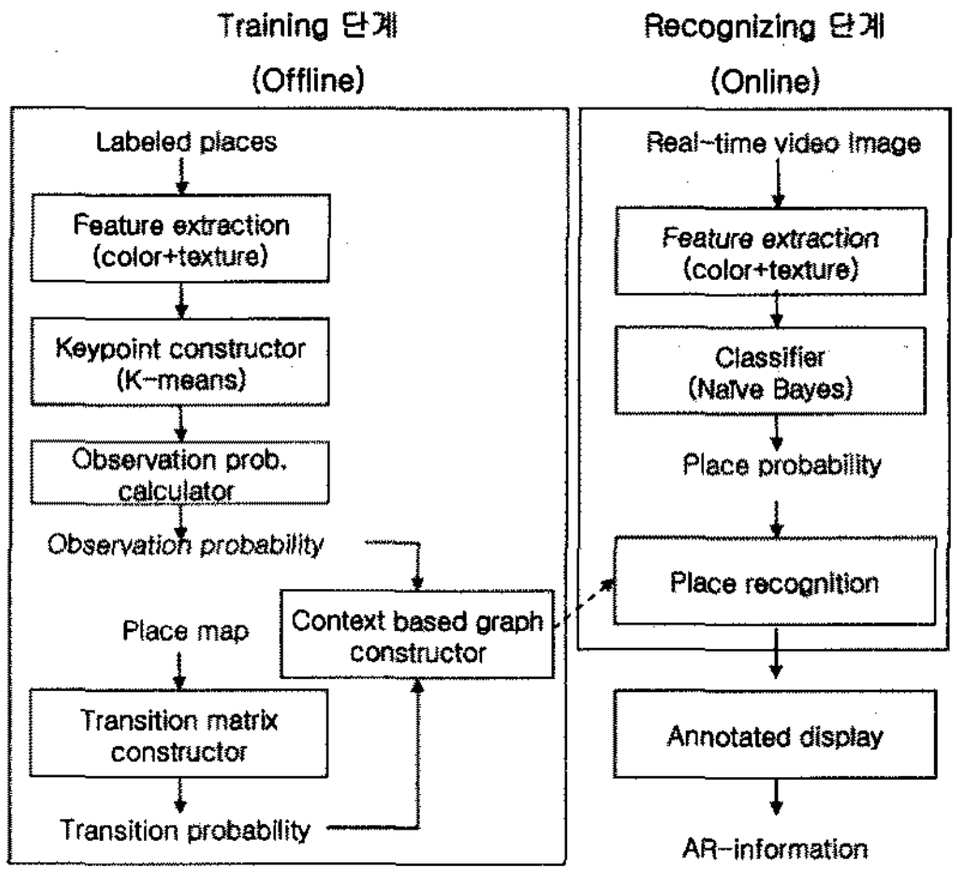

그림 1 전체 구조도

실시간 인식단계에서는 매순간의 비디오 프레임으로부터 특징을 추출한 후, 학습단계에서 구축된 컨텍스트 그래프 모델을 이용해 각 장소별 확률 중 가장 높은 확률에 해당하는 장소를 선택한다. 이 결과에 따라 모바일 장치에 사용자 문맥에 맞는 정보를 디스플레이할 수 있게 된다.

3.1 장소인식을 위한 그래프 기반 모델

영상인식을 보완할 장소문맥 정보를 활용하기 위하여 그래프 기반 모델인 컨텍스트 그래프를 이용하였다. 이 모델에서 상태는 인식하고자 하는 사용자의 장소에 해당하고 관측값은 카메라로부터의 영상신호가 된다. 한 장소에서 다른 장소로의 이동확률은 장소간의 인접관계를 그래프로 나타낸 위치관계도(place-map)로부터 계산된다. 그림 2의 상단부는 컨텍스트 그래프이며 하단부는 위치관계도(place-map)이다.

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 $P(Q=q|Z_{1:t})$이고, 다음과 같이 계산될 수 있다.

$$P(Q_t=q|Z_{1:t}) \propto p(Z_t|Q_t=q)P(Q_t=q|Z_{1:t-1}) \quad (1)$$

where

$$P(Q_t=q|Z_{1:t-1}) = \Sigma P(Q_t=q|Q_{1:t-1}) P(Q_{1:t-1}|Z_{1:t-1})$$

$P(Z_t|Q_t=q)$는 주어진 장소에서 이미지 특징벡터가 관측될 확률을 의미하고, $P(Q_t=q|Q_{1:t-1})$는 이전 시간까지의 장소에서 현재시간의 장소로 이동할 확률을 의미한다. 식 (1)과 같이 $P(Q_{1:t-1}|Z_{1:t-1})$은 관측확률과 장소이동확률을 이용하여 재귀적으로 계산할 수 있다.

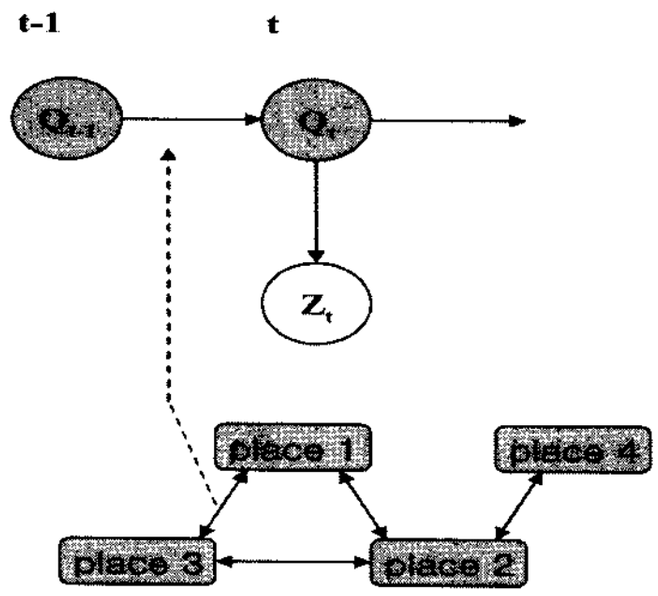

그림 2 컨텍스트 그래프와 위치관계도

3.2 이미지 특징 추출

실시간 모바일환경에서 장소인식이 구현되기 위해서는 계산량이 적으면서 영상 구분능력이 뛰어난 특징 추출방식이 요구된다. 본 논문에서는 전역적 특징인 칼라와 텍스쳐 정보를 활용하면서, 장면의 부분적 폐색에 민감한 전역적 특징들의 단점을 보완하기 위하여 분할된 영상 조각들의 특징을 구함으로써 지역적 특징도 반영하였다. 먼저, 장면 이미지를 10×10 고정 격자로 나누어 지역 패치로 구분한다. 각 패치를 묘사하는 특징벡터는 두 종류의 값으로 구성되는데, 하나는 HSI 칼라 히스토그램, 또 하나는 텍스쳐 특징값이다. 칼라는 색조와 명도만을 고려한다. 장면 이미지에 마스크를 씌워 밝기값으로 구한 픽셀의 공분산 행렬에서, 3개의 텍스쳐 특징값을 계산한다. 이미지의 특징벡터는 한 패치당 25차원으로 계산해 사용하였다[8]. 그림 3은 이미지로부터 지역 패치를 적용하여 특징을 추출하는 과정을 나타낸 것이다.

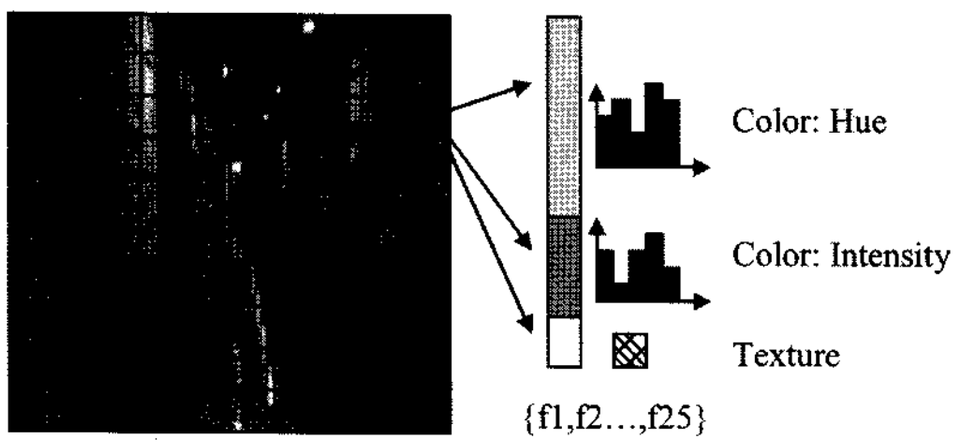

그림 3 지역 패치를 이용한 특징 추출

3.2.1 HSI 칼라 히스토그램

칼라를 표현하는데 있어서 RGB색상체계는 컴퓨터 그래픽에서 널리 사용되지만, 인간의 칼라인식은 RGB보다는 HSI/HSV(Hue, Saturation, and Intensity / Value) 칼라공간에 더 유사하다. 따라서 HSI 칼라공간은 컴퓨터비전에서 카메라영상의 잡음이나 조명에 좀더 강건한 색체계로 주로 사용이 된다. 이러한 이유로 본 논문에서는 HSI 칼라공간을 사용하였다.

사람의 시각이 실제로 인식할 수 있는 칼라범주는 한정되어있는데, 대부분의 연구에서는 대략 11개~30가지의 한정된 분류를 이용하였다[8]. 본 논문에서는 색조공간을 16개의 세그먼트로, 명도공간을 6개의 세그먼트로 나눠서 히스토그램으로 표현하였다.

x값을 가지는 픽셀 j에 대해 빈(bin) b_i는 다음과 같은 수식을 이용하여 결정된다.

$$b_i = \frac{x_j}{s} \tag{2}$$

여기서 빈 사이의 간격 s는 다음과 같다.

$$s = \frac{\max(x) - \min(x)}{n} \tag{3}$$

$\max(x)$와 $\min(x)$은 각각 해당 빈의 픽셀 x값의 최대, 최소값이다. 식 (2)과 (3)에 의해 다음과 같이 쓸 수 있다.

$$b_i = \frac{n \cdot x_j}{\max(x) - \min(x)} \tag{4}$$

각 픽셀의 분포횟수는 이산화(discretization)로 인한 에러에 대처하기 위해 한 픽셀 당 3개의 빈에 다음과 같은 비율로 나누어 누적된다.

$$\begin{aligned} b_i &\quad += 1 \\ b_{i-1} &+= 1 - \frac{|x_j - \min(b_i)|}{\max(b_i) - \min(b_i)} \\ b_{i+1} &+= 1 - \frac{|x_j - \max(b_i)|}{\max(b_i) - \min(b_i)} \end{aligned} \tag{5}$$

$$where\ \min(b_i) \le x_j \le \max(b_i),$$
$$with\ i \in \{0, n-1\}\ \text{and}\ j \in \{0, p-1\}$$

3.2.2 텍스쳐 성질

텍스쳐 분석은 대부분 픽셀의 칼라보다는 명도를 주로 이용한다. 본 논문에서는 N×N 마스크를 이용하여 마스크내 픽셀 명도의 공분산 행렬을 구한 후, 공분산과 고유값으로부터 3개의 특징값을 계산하였다. 여기서 마스크의 사이즈는 전역적, 지역적인 특징을 얼만큼의 비율로 고려하냐에 따라 달라질 수 있다.

각 마스크의 픽셀에서 각각 수평적이고 수직적으로 놓여있는 주변픽셀들의 차이를 통해 텍스쳐의 성질을 반영할 수 있도록 식 (6)과 같이 공분산 행렬을 구한다. 이 때 함수 f는 정규화된 픽셀 (i,j) 값이다.

각 마스크에 대하여 공분산 행렬 M이 구해지면, 공분산 행렬의 고유값 $ev1, ev2$를 구한 후, 식 (7)을 통하여 3개의 텍스쳐 특징값을 얻게 된다.

$$\begin{aligned}&\text{for } each(p,q) \in Image \\ &\quad \text{for } each(i,j) \in Mask(p,q) \\ &\qquad Sum += f(i,j) \\ &\qquad SqSum += f(i,j)^2 \\ &\qquad M_{11} += (f(i+1,j)-f(i-1,j))^2 \\ &\qquad M_{12} += (f(i,j+1)-f(i,j-1))^2 \\ &\qquad M_{22} += f(i+1,j)-f(i-1,j))* \\ &\qquad\qquad (f(i,j+1)-f(i,j-1))\end{aligned} \tag{6}$$

$$F_1 = SqSum - Sum^2 \tag{7}$$

$$F_2 = \frac{\min\{ev_1, ev_2\}}{\max\{ev_1, ev_2\}}$$

$$F_3 = \max\{ev_1, ev_2\}$$

분산값인 F_1은 이미지에서 텍스쳐의 전역적인 특징을 표현하게 되고, F_2와 F_3는 텍스쳐의 구조를 표현하게 되어, 이러한 세 값의 조합은 결과적으로 이미지의 전역적이면서 지역적인 성질을 반영하는 특징값이 된다. 여기서 F_2과 F_3는 회전불변(rotation-invariant) 특성을 지닌다. 이러한 방법은 분할된 패치에 속한 픽셀들의 텍스쳐 특징을 통계적으로 반영하므로 특징 벡터 값이 좀 다르게 산출되더라도 비슷한 벡터로 군집화될 수 있다. 또한, 실시간에 주어지는 영상이 관심 장소의 중심에서 옆으로 약간 이동(shift) 되었더라도 학습 데이타 자체가 그러한 변동을 흡수할 수 있게끔 관심 장소 주변 영상들을 충분히 좁은 간격으로 샘플링하여 수집되므로 영상 쉬프트로 인한 오인식 문제를 피할 수 있다.

3.3 장면 분류 및 관측확률 추정

실시간으로 영상에서 추출된 특징에 가장 잘 정합되는 장소를 파악하기 위해서는 학습단계에서 관측 확률(Observation Probability)를 계산한 후, 컨텍스트 그래프 모델의 노드(장소)에 저장해 두어야한다. 관측확률은 표본 데이타로부터의 학습을 통해서 추정되는데, 이는 레이블링된 학습 데이타를 장소에 따라 분류시킴으로써 계산할 수 있다. 본 장소인식 시스템에서는 이를 위해 백 오프 키포인트 방식(Bags of Keypoints method)을 이용한다[10].

백 오프 키포인트 방식(Bags of Keypoints method)은 특징벡터 공간에서 군집화(그림 4 참조, K-means 활용)된 벡터들의 중점인 키포인트를 구한 후, 키포인트 누적 히스토그램을 이용하여 영상을 분류하는 방법이다. 즉, 분류 결과는 그림 5와 같이 주어진 영상에서 특정 패턴이 나타나는 횟수를 히스토그램으로 만든 것이다. 구현시에 사용되는 분류기로는 Naïve Bayes방식과 SVM방식을 이용하는데, Naïve Bayes 분류기가 장소인식 시스템에 적용하기에 더 간단하고 계산량이 작아 효율적이므로 본 논문에서는 Naïve Bayes 분류기를 사용하였다.

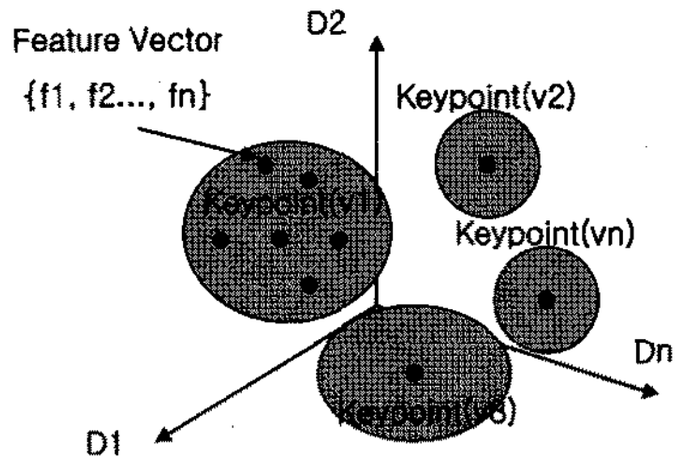

그림 4 K-means를 이용한 특징벡터 군집화 및 키포인트 추출

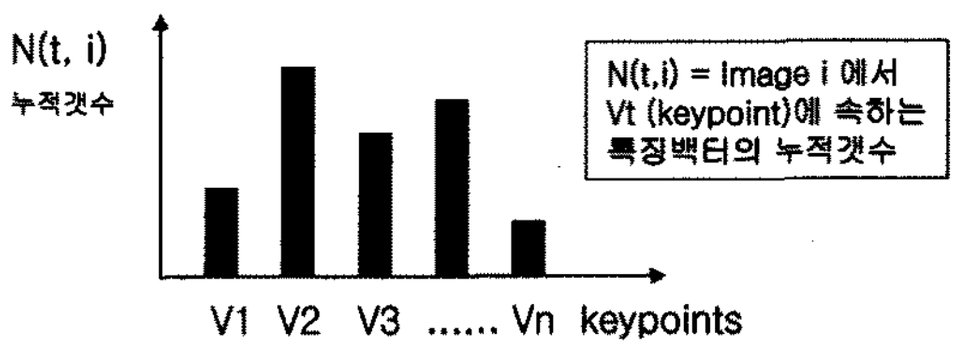

그림 5 키포인트 누적 히스토그램(Bag of keypoints)

즉, 학습단계에서 $I=\{I_i\}$는 레이블링된 영상들의 집합이고, $V=\{V_i\}$는 대표 키포인트들이라고 할 때, 영상으로부터 추출된 특징벡터들은 특징벡터 공간에서 가장 가까운 키포인트에 귀속되고 각 프레임 I_i에서 키포인트 V_t에 누적되는 특징벡터의 수를 카운트하여 N(t,i)라고 한다. 이에 대해 실시간에는 입력되는 새로운 영상을 분류하기 위하여 Bayes 규칙을 적용하여 가장 큰 사후 점수(posteriori score)를 얻음으로써 장소를 인식한다.

$$P(C_j|I_i) \propto P(C_j)P(I_i|C_j) = P(C_j)\prod_{t=1}^{|V|} P(V_t|C_j)^{N(t,i)} \tag{9}$$

$$P(V_t|C_j) = \frac{1+\sum_{\{I_i=C_j\}} N(t,i)}{|V|+\sum_{s=1}^{V}\sum_{\{I_i=C_j\}} N(s,i)}$$

여기서 $P(I_i|C_j)$는 장소인식 시스템에서 $P(Z_t|Q_t \equiv q)$에 해당하는데, 특정한 위치(장소)에서 특정한 영상 특징패턴이 나타날 확률이다. 레이블링된 이미지 집합으로부터 N(t,i)를 모두 구한 후, 새로운 이미지 I_i와 특정 장소 C_j를 입력값으로 하여 관측확률(Observation Probability)를 구할 수 있게 된다.

4. 구현 및 결과

제안된 방법은 P4 1.7Ghz, 512M RAM 조건의 UMPC에서 구현되었으며, 이화여대 SK텔레콤관 건물의 각 층을 이동하면서 세부 장소를 인식하고 정보를 증강하는 시스템을 구축하여 실험하였다. 총 12곳의 장

소를 대상으로 레이블링된 이미지를 학습데이타로 이용하였으며, 훈련 이미지는 90 프레임을 사용하였다. 각 장소별로 7~8 프레임 정도가 사용된 셈이다.

실시간 테스트에서는 320*240 크기의 연속된 비디오 프레임 각각을 12곳의 장소 중 하나로 분류하였으며, 그 결과, 표 1과 같이 본 시스템의 프로세싱 타임은 약 20 f/s에 이르며, 약 72~89% 정도의 인식률을 보였다. 약 70~90%의 인식률을 보였던 기존 연구[2-7]들에 비해 인식률의 뚜렷한 향상은 없지만, 훈련 데이타가 장소 당 7~8개 정도였음을 감안할 때 좋은 성능이라고 할 수 있다. 또한, 이들 연구들이 2~4 f/s에 해당하는 느린 속도를 보이는 것에 비해 본 연구는 실시간성 면에서 뛰어난 성능을 보인다. 이러한 실시간성은 즉각적 피드백을 필요로 하는 이동형 시스템에서 필수적인 것이다.

표 1 속도 및 인식률 실험 결과

Frame/sec	인식률	키포인트 개수	해상도
약 20 f/s	약72~89%	100	320x240

그림 6은 구현된 시스템에서 학습 및 인식률 테스트 인터페이스를 보여주고 있고, 그림 7은 실시간 이동형 시스템에 적용한 정보 증강 화면이다.

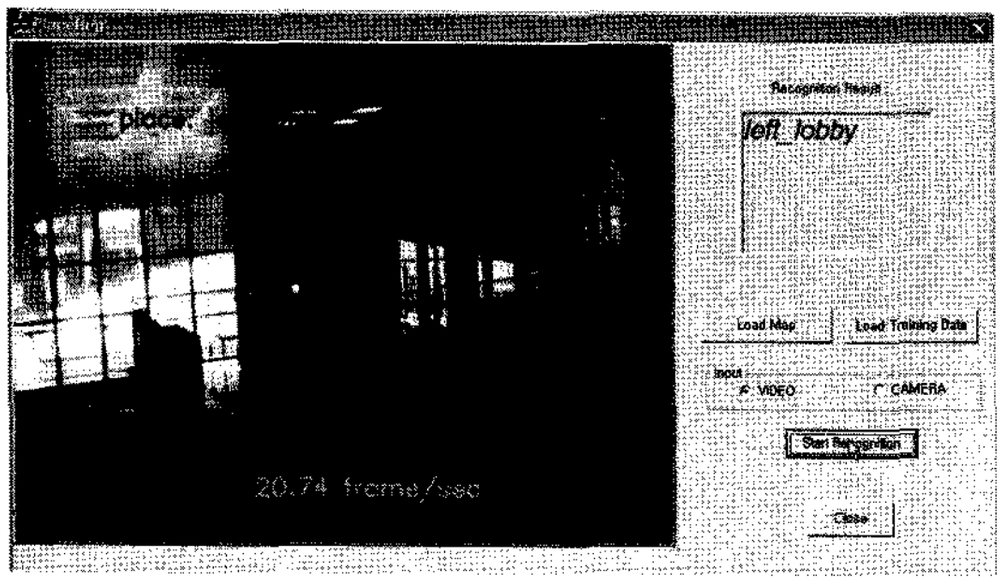

그림 6 장소 인식 시스템 테스트베드

그림 7 이동형 정보 증강 화면

5. 결론 및 향후연구

본 논문에서는 모바일 혼합/증강현실에 적용하기 위하여 계산량이 많은 특징추출방법을 이용했던 기존의 연구들과 달리, 모바일컴퓨팅에 적합한 장소인식을 위해서 단일카메라를 이용하고 실시간성을 높이는 알고리즘을 제안하였다. 실험결과, 전체 인식률을 유지하면서 비교적 간단한 특징추출계산을 통해 실시간성이 높아짐을 확인할 수 있었다. 향후 연구에서는 외부조명상태에 좀 더 강건하도록 장소인식 알고리즘을 보완할 필요가 있다.

참 고 문 헌

[1] A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *In IEEE International Conference on Computer Vision*, Vol.1, pp. 273-280, Nice, France, 2003.

[2] D. Jung, H. Kim, "Place Recognition System from Long-Term Observations," *Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, IEA/AIE, pp. 36-43, 2005.

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

[4] 황금성, 조성배, "불확실한 장면의 효과적인 인식을 위한 베이지안 네트워크의 온톨로지 기반 제한 학습방법", *정보과학회논문지: 소프트웨어 및 응용*, 제34권, 6호, pp. 100-109, 2007.

[5] S. Kim, I. Kweon, "Collaborative Place and Object Recognition in Video using Bidirectional Context Information," *제1회 한국지능로봇 하계종합 학술대회 논문집*, pp. 432-439, 2006.

[6] S. Im, S. Cho, "Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform," *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pp. 1080-1087, Antwerp, Belgium, Sept. pp. 18-21, 2006.

[7] 민경민, 이성훈, 김동호, 김진형, "Nonstationary HMM을 이용한 다중 카메라 기반 장소 인식", *HCI 2007*.

[8] 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, pp. 38-47, 2004.

[9] L. FeiFei, P. Perona. "A Bayesian Hierarchical Model for Learning Natural Scene Categories," IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2005.

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