네비게이션을 위한 문자영상기반의 영상매칭 방법

박안진[O] 정기철
숭실대학교, 정보과학대학, 미디어학과, HCI Lab.
{anjin[O], kcjung}@ssu.ac.kr

Text Cues-based Image Matching Method for Navigation

Anjin Park[0] Keechul Jung
HCI Lab., School of Media, College of Information Science, Soongsil University.

요 약

유비쿼터스 시대가 다가오면서, 많은 사람들은 모르는 장소에서 자신의 위치와 목적지까지의 경로에 대한 정보를 알고 싶어할 것이다. 기존의 네비게이션(navigation)을 위한 비전기술은 고차원과 저차원 특징값을 이용하였다. 텍스춰 정보, 색상 히스토그램과 같은 저차원 특징값은 영상의 특징을 정확하게 표현하기 어려우며, 마커와 같은 고차원 정보는 실험환경을 구축하는데 어려움이 있다. 우리는 기존 저/고차원의 특징값 대신, 영상의 특징을 표현하고 인덱싱(indexing)하기 위한 유용한 정보를 많이 포함하고 있으며, 실제환경에서 널리 분포되어있는 중차원 특징값인 문자영상을 이용한다. 문자영상추출은 MLP(multi-layer perceptron)와 CAMShift알고리즘을 결합한 방법을 이용하며, 서로 다른 장소지만 같은 문자를 가진 곳에서 인식을 수행하기 위해 문자영상의 크기와 기울기를 기반으로 한 영상 검색공간을 대상으로 영상매칭을 수행한다. 실험에서 문자영상을 포함하는 직사각형 검색공간으로 인해 다양한 크기와 기울기에서 높은 인식률을 보이며, 간단한 계산으로 빠른 수행시간을 가진다.

1. 서 론

유비쿼터스 시대가 다가오면서, 가까운 미래에는 거의 대부분의 사람들이 모바일 장치를 이용할 것이며, 일상생활에서 널리 이용될 것이다. 이런 시대에서 많은 사람들은 모르는 장소에서 자신의 위치와 목적지까지의 경로를 모바일 장치를 통하여 알고 싶어할 것이다.

실외에서 사용자의 위치를 알려주는 GPS는 실내에서는 작동이 되지 않는 단점을 가지고 있다. 이를 위해 실내에서 상대적인 좌표를 구하기 위해 사용되는 각종 센서들은, 작동을 위해 배터리가 필요하며 설치와 운영을 위해 많은 비용이 필요한 문제점을 가지고 있다[1].

그래서, 가격이 저렴한 비전 기반의 네비게이션(navigation) 시스템에 대한 많은 연구가 진행되어 왔다[2-6]. Aoki[2] 등은 웨어러블(wearable) 컴퓨터에서 입력받은 연속영상에서의 장소인식에 대한 연구를 수행하였으며, 밝기 변화에 상대적으로 영향을 덜 받는 휴(hue) 히스토그램과 DTW를 이용하였다. 이 방법은 장소는 다르지만 유사한 휴값을 가진 곳에서 인식률이 떨어지는 단점이 있다. 그리고, DB에 있는 장소 영상과 입력 받은 정지 영상을 매칭(matching)하여 장소를 인식하는 연구도 진행중이다[3-4]. Wolf[3] 등은 Monte-Carlo 방법을 이용한 영상 검색 시스템을 기반으로 비전 기반의 모바일 로봇 위치 시스템을 제안하였으며, Kosecka와 Yang[4]은 특정 검색공간의 크기에 강건한 특징값을 이용하여 장소를 인식하는 방법을 제안하였다. 그러나 DB를 이용한 장소 인식 방법은 대략 2~3 미터 단위로 영상을 DB에 저장하고 있어야 하며, 한 건물내의 모든 위치를 기억하기 위해 많은 영상이 필요하다. DB를 쉽게 획득하기 위해, Kourogi[5] 등은 파노라마 영상을 이용하였으며, 인식을 위해 영상 레지스트레이션(registration) 방법을 이용하였다. 앞의 세 방법[3-5]은 영상 매칭과 레지스트레이션을 위해 저차원의 특징값을 이용하며, 복잡한 계산으로 인해 많은 수행시간과 낮은 인식률을 보인다. 비전 기술을 이용하여 정확한 위치를 인식하는 또 다른 연구로써 실제환경에 설치된 인위적인 마커를 이용하였으며[6], 이 방법은 수동적으로 마커를 설치해야 하는 단점을 가진다.

본 논문에서는 저차원의 특징값대신 중차원 특징값인 문자영상을 이용하여 파노라마 영상과 입력 영상의 영상 매칭 방법을 제안한다. 문자영상은 영상의 특징을 표현하고 인덱싱(indexing)하기 위한 유용한 정보를 많이 포함하고 있으며, 얼굴, 사람 몸과 같은 다른 의미있는 대상(sementaic objects)보다 쉽게 추출할 수 있는 장점을 가지고 있다[7]. 기존 문자인식 후 장소를 인식하는 방법[8]은 서로 다른 장소에서 같은 문자를 가진 경우 모호한 인식결과를 보이는 단점이 있으며, 우리는 문자영역의 크기와 기울기를 기반으로 한 직사각형 검색공간을 이용하여 이를 해결한다. 실험결과에서 우리의 방법은 저차원의 모호함[2]과 복잡한 계산[3-5]없이 중차원의 특징값만으로 매칭을 수행하므로 높은 인식률과 빠른 수행시간을 보이며, 문자영상은 건물내의 어디에서나 분포되어 있기 때문에, 고차원 특징값의 단점인 미리 설치되어 있는 장소에서만 수행이 가능한 단점을 해결한다.

본 논문의 구성은 다음과 같다. 제 2장에서 우리시스템의 전체 흐름도를 보여주며, 제 3장에서 문자 추출방법, 제 4장에서 DB의 파노라마 영상과 입력 영상의 문자 정보 사이의 유사도 측정 방법을 기술한다. 제 5장에서는 실험결과에 대해 서술하며, 제 6장에서 향후 연구방향을 기술한다.

2. 전체 흐름도

우리는 모바일 장치에서 입력된 영상의 문자영상과 파노라마 영상의 문자영상들간의 유사도를 이용하여 영상 매칭을 수행하며,

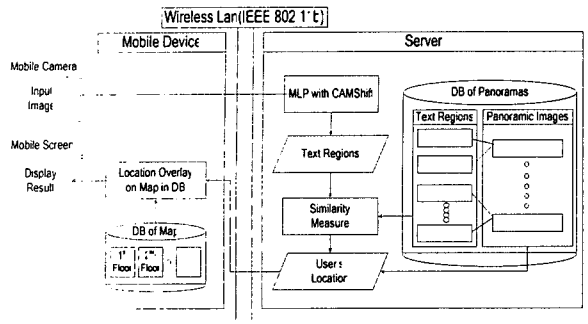


그림 1. 제안된 방법의 전체 흐름도.

이를 기반으로 네비게이션 시스템을 제공한다(그림 1).

모바일 장치는 부족한 연산자원, 제한된 저장공간과 같은 문제점[9]을 가지고 있기 때문에, 우리는 문자영역추출, 유사도측정과 같은 복잡한 연산 수행과, 고용량의 DB 영상을 위해 서버(server)를 이용한다. 문자영상과 중심값, 크기, 기울기와 같은 특징값을 추출하기 위해 MLP(multi-layer perceptron)와 CAMShift 알고리즘을 결합한 문자영역추출방법[7]을 이용하며, DB를 쉽게 구축하기 위해 파노라마 영상[3]을 이용한다. 같은 문자영상을 가지지만 장소가 다른 곳을 인식하기 위해 문자영상을 포함하는 검색공간을 이용하며, 모바일 장치에서 비전 시스템을 수행하기 위해 밝기 변화나 잡음에 강건해야 하기 때문에, DB와 입력영상의 유사도는 검색공간내의 휴 히스토그램을 이용하여 측정한다[2]. 입력영상의 문자영역을 DB와 빠르게 비교하기 위해 파노라마 영상에서의 문자영역을 포함하는 검색공간을 따로 저장하며, 검색공간은 문자영상의 특징값을 기반으로 한 직사각형으로 한다. 파노라마 영상은 지도상의 위치를 가지고 있으며, 우리는 매칭된 결과를 이용하여 사용자에게 자신의 위치가 표시된 지도를 제공한다.

3. 문자영역추출

우리는 문자추출을 빠르고 효과적으로 수행하기 위해 MLP와 CAMShift 알고리즘[7]을 이용한다.

본 논문에서는 MLP를 이용하여 다양한 크기나 모양의 문자와 배경에 적응할 수 있는 텍스춰(texture) 분석기를 구성한다. MLP는 2개의 은닉층, 1개의 출력노드로 구성되며 인접층의 노드들은 모두 연결되어 있고, 입력층은 그림 2와 같이, 입력 영상에 M×M 크기의 입력창 내의 검은색으로 표시된 화소들의 인텐시티 값을 사용한다. 이러한 입력 구조는 텍스춰 분석 분야에서 성능과 속도 향상에 좋은 것으로 알려져 있다[10]. MLP를 이용하여 입력 영상을 스캔함으로써 생성된 결과 영상은 입력 영상의 각 화소를 문자와 비-문자 클래스로 구분한다.

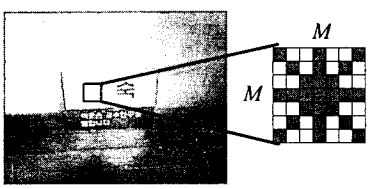


그림 2. 신경망의 입력층.

기존 텍스춰 분류 방법은 입력 영상의 전체 영역을 탐색하여 많은 수행 시간을 가지는 단점이 있다. 우리는 이런 문제를 해결하기 위해 변형된 CAMShift 알고리즘을 사용 한다.

본 논문에서는 TPI(text probability image)상에서 CAMShift 알고리즘을 반복 수행함으로써 영상 내의 문자를 검출한다. 시작단계에서 TPI상의 각 검색창이 문자 영역을 포함하고 있는지를 결정하며, 연속된 일련의 단계에서 2차원 모멘트 계산을 통해서 문자 영역의 크기와 위치를 구하고 인접한 픽셀을 병합하면서 문자 영역을 찾게 된다. 매 번의 반복 수행과정에서 검색창의크기를 문자 영역의 크기에 비례해서 수정하고, 겹치는 노드들을 제거하기 위해 노드 병합과정을 수행한다. 수렴된 후 각 문자 영역들은 크기와 가로 세로 비율을 이용하여 필터링을 거친다.

문자 영역의 위치와 크기를 구하기 위해 연산이 간단하고 잡음에 효과적인 모멘트를 사용한다. 이차원 p+q차 모멘트는 다음과 같이 기술할 수 있다.

$$M_{pq}(i) = \sum_x \sum_y x^p y^q TPI(x,y) \quad (1)$$

문자 영역의 중심 좌표는

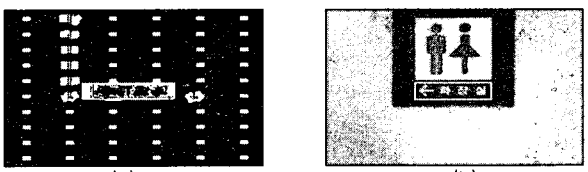
$$mean_x(i)_t = M_{10}/M_{00}, mean_y(i)_t = M_{01}/M_{00} \quad (2)$$

로 설정할 수 있고, 문자 영역의 폭과 높이, 기울기는 다음과 표현할 수 있다.

$$width(i)_t = \sqrt{2(a+c)+2\sqrt{b^2+(a-c)^2}},$$
$$height(i)_t = \sqrt{2(a+c)-2\sqrt{b^2+(a-c)^2}},$$
$$\theta(i)_t = \arctan(2b/(a-c))/2 \quad (3)$$

where $a = \frac{M_{20}}{M_{00}} - \left(\frac{M_{10}}{M_{00}}\right)^2, b = 2\left(\frac{M_{11}}{M_{00}} - \frac{M_{10}M_{01}}{M_{00}^2}\right), c = \frac{M_{02}}{M_{00}} - \left(\frac{M_{01}}{M_{00}}\right)^2$.

CAMShift의 수행시간을 줄이기 위해, 노드들의 겹쳐진 비율이 일정 이상의 두 노드를 병합한다. 그리고, 문자 영역의 특징값이 더 이상 변화하지 않을 때까지 반복한다. 그림 3은 MLP와 CAMShift를 이용한 문자추출결과를 보여준다. 그림 3(a)에서 검은색 화소들은 MLP에 의해 문자 영역으로 표시된 부분이며, 회색 부분은 MLP가 수행되지 않는 영역들이다.

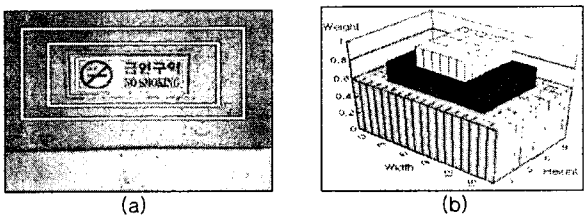


(a) (b)

그림 3. 문자추출결과: (a) 신경망 결과영상, (b) 마킹된 결과영상.

4. 유사도 측정

우리는 문자영상을 포함하는 검색공간내의 휴 히스토그램을 이용하여 유사도를 측정한다. 검색공간의 모양은 각 문자영상마다 가로, 세로의 길이가 다양하기 때문에, 가로, 세로를 기반으로 직사각형으로 하며, 문자영상의 원근에 의한 왜곡, 기울어짐과 같은 여러 환경에서 유사도를 측정하기 위해, 검색공간을 3 단계로 나누어 중심에 더욱 가중치(weight)를 두며(그림 4), 히스토그램을 만들 때 가중치를 적용한다.



(a) (b)

그림 4. 검색공간: (a) 3단계의 검색공간, (b) 검색공간 가중치.

검색공간의 다양한 크기에 적용하기 위해 히스토그램 값을 검색 공간 픽셀 수로 정규화(normalize)시키며, 계산량을 줄이기 위해 히스토그램의 빈수를 36으로 줄인다. 그리고, DB와 입력 영상의 히스토그램의 유사도를 측정하기 위해 유클리디언 거리(Euclidean distance)를 이용한다(그림 5).

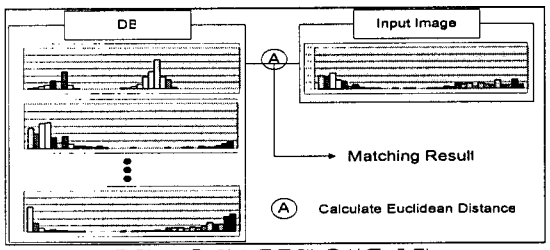


그림 5. 휴 히스토그램 유사도 측정.

5. 실험 및 결과

우리는 클라이언트/서버 구조를 이용하여 모바일 기반의 네비게이션 시스템을 구현하였으며, 모바일 장치는 Pocket PC-2003기반의 POZ x301을 이용하였으며, 카메라는 30만 화소이다.

그림 6은 검색공간 내의 휴 히스토그램을 보여준다. 그림 6(a)의 내부 직사각형은 문자영상이며, 외부 직사각형은 검색공간이다. 우리는 검색공간의 크기로 DB의 파노라마 영상에서는 문자영상의 가로, 세로길이의 5배로 하며, 입력영상에서는 입력영상의 제한된 크기를 고려하여 가로, 세로길이를 선택한다. 그리고, 유사도를 측정할 때 입력영상의 검색공간 크기에 맞추어 DB의 검색공간의 크기를 변형한다. 그림 6(a)의 검색공간은 문자영상의 가로, 세로길이의 두 배이다. 그림 6에서 보는 바와 같이 같은 문자영상이라도, 장소가 다르기 때문에, 다른 휴 히스토그램을 보인다.

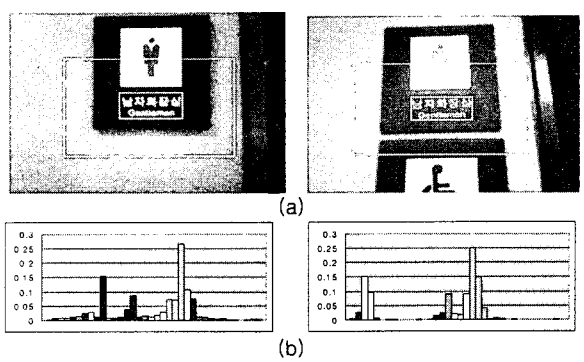


그림 6. 검색창 내의 휴 히스토그램: (a) 입력영상, (b) 휴 히스토그램.

실험에서 검색공간에 대한 두 가지 문제점을 가지고 있다. 첫째 DB의 파노라마 영상에서 문자영상이 외각에 있을 경우이며, 검색공간크기 때문에 검색공간 내에 파노라마 영상이 존재하지 않는 문제가 생긴다. 둘째 입력영상에서 문자영상을 크게 입력받아 검색공간과 문자영상이 같은 경우이며, 기존 문자인식 후 장소를 인식하는 방법과 같은 문제점을 가진다. 본 논문에서는 두 가지 문제점을 제한하고 실험을 수행하였다.

표 1은 각 단계별 수행시간(ms)을 보여준다. PDA에서 PC로는 입력 영상을 넘겨주며, 문자추출(MLP with CAMShift)과 인식을 수행한 뒤 맵 인덱스와 맵에서의 위치를 PC에서 PDA로 전송한다. 우리의 방법은 문자 영상을 추출하기 위해 약 550ms의 수행시간을 보이며, 초기 5~10초의 시간이 필요하고, 수행시 700~800ms가 필요한 영상 정합 방법[3]보다 빠른 수행시간을 보인다.

표 1. 각 단계별 수행시간(ms)

	PDA →PC	문자추출	인식	PC→ PDA	Total
수행시간	130	300	100	20	550

6. 결론

우리는 텍스춰와 같은 저차원 특징값, 인공적인 마커와 같은 고차원 특징값 대신 중차원 특징값인 문자영상을 이용하여 파노라마 영상에서의 영상매칭 방법을 제안했다. 문자영상은 영상을 표현하기 위한 유용한 정보를 가지고 있으며, 실제환경에 널리 분포되어 있는 장점을 가지고 있어, 실내에서 네비게이션 시스템을 구축할 때 많은 장점을 가지고 있다.

우리 방법의 가장 큰 단점은 건물 내 문자가 없는 지역에서는 수행이 되지 않는 단점을 가지고 있으며, 이런 장소에서의 문제점을 보안하기 위해 문자영상과 연계한 또 다른 효과적인 특징값을 이용하여 더욱 정확한 네비게이션을 구축할 것이다.

7. 참고문헌

[1] D. L. Ipina, P. R. S. Mendonca and A. Hopper, "TRIP: a Low-Cost Vision-based Location System for Ubiquitous Computing," Personal and Ubiquitous Computing, Vol. 6, Issue 3, pp. 206-219, May 2002.

[2] H. Aoki, B. Schiele and A. Pentland, "Realtime Personal Positioning System for a Wearable Computers,"pp. 37-43, Oct. 1999.

[3] J. Wolf, W. Burgard and H. Burkhard, "Robust Vision-based Localization for Mobile Robots using an Image Retrieval System based on Invariant Features," Proceedings of International Conferences on Robotics and Automation, pp. 359-365, May 2002.

[4] J. Kosecka and X. Yang, "Location Recognition and Global Localization based on Scale Invariant Features," Proceedings of European Conference on Computer Vision, 2003.

[5] M. Kourogi, T. Kurata and K. Sakaue, "A Panorama-based Method of Personal Positioning and Orientation and Its Real-time Applications for wearable Computers," Proceedings of International Workshop of Wearable Computers, pp. 107-114, 2001.

[6] W. Piekaraski, B. Avery, B. H. Thomas and P. Malbezin, "Integrated Head and Hand Tracking for Indoor and Outdoor Augmented Reality," Proceedings of IEEE Virtual Reality, pp. 267-271, Mar. 2004.

[7] K. Jung, K. I. Kim and A. K. Jain, "Text Information Extraction in Images and Videos: a Survey," Pattern Recognition, Vol. 37, Issue 5, pp. 977-997, May 2004.

[8] Y. Liu and T. Yamamura, "Character-based Mobile Robot Navigation," Proceedings of International Conference on Intelligent Robots and Systems, No. 2, pp. 610-616, 1999.

[9] A. Park and K. Jung, "PDA-based Text Localization System using Client/Server Architecture," Proceedings of Pacific Rim International Conference on Artificial Intelligence, Lecture Notes on Computer Sciences, Vol. 3175, Aug. 2004.

[10] A. K. Jain, K. Karu, "Learning Texture Discrimination Masks," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 2, pp. 195-205, 1996.