동적 베이지안 네트워크를 이용한 컨텍스트 기반 장소 및 물체 인식

임승빈° 조성배
연세대학교 컴퓨터과학과
envymask@sclab.yonsei.ac.kr°, sbcho@cs.yonsei.ac.kr

Context based Place and Object Recognition using Dynamic Bayesian Network

Seung-bin Im°, Sung-bae Cho
Dept. of Computer Science, Yonsei University

요 약

영상 이해는 컴퓨터 비전의 가장 높은 수준의 처리 기법이다. 영상을 이해하기 위해서는 위치 정보, 물체 존재정보와 같은 기본 컨텍스트들을 추출하는 것이 중요하다. 그러나 실내 환경의 영상 정보는 카메라의 흔들림이나 각도, 빛의 상태에 따라 불확실해지기 때문에 이러한 불확실함에 강인한 영상 인식 기법이 필요하다. 동적 베이지안 네트워크(DBN)는 이러한 불확실한 정보의 처리에 강인하며 장소와 물체의 관계 등 고수준의 컨텍스트를 모델링하는데 좋은 성능을 보이는 확률 모델이다. 또한 DBN은 이전 상태를 추론에 활용할 수 있으므로 장소 인식과 같은 컨텍스트의 추출에 좋다. 본 연구에서는 불확실한 실내 환경 영상으로부터 영상 전처리를 통해 특징값을 추출하고, 회전이나 크기 변화에 강인한 물체인식기법인 크기 불변 특징 변환기법(SIFT)을 이용하여 물체 존재정보를 추출하여 고수준 컨텍스트가 모델링된 DBN 추론으로 장소 및 물체를 인식하는 방법을 제안한다. 실제 대학 실내 환경에서의 실험으로 DBN을 이용한 영상 인식기법이 좋은 성능을 보임을 확인할 수 있었다.

1. 서 론

영상 이해는 지능 로봇 분야에서 아직 해결하지 못한 어려운 문제 중 하나이다. 영상을 이해하기 위해서는 먼저 영상을 구성하고 있는 기본적인 정보들을 추출하고, 추출된 정보들을 기반으로 고수준의 추론을 수행하여야 한다. 영상은 의미 있는 다양한 정보들로 구성되어 있다. 따라서 영상을 이해하기 위해서는 이러한 컨텍스트들의 인식과 그것들 사이의 관계에 대한 이해가 필요하다. 그러므로 장소 정보나 물체 존재정보 같은 기본적인 컨텍스트의 추출로부터 영상 이해에 접근하는 것은 고수준의 영상 이해를 풀기 위한 좋은 접근방식이 된다.

영상은 많은 양의 정보를 담고 있지만 카메라의 흔들림이나 각도, 빛의 상태에 따라 불확실해진다. 그러므로 영상 이해를 위한 컨텍스트의 추출은 이러한 불확실성을 해결할 수 있어야 한다. 장소와 그 장소에 존재하는 물체들의 관계, 또는 어떤 활동과 그에 필요한 물체사이의 관계와 같은 고수준의 컨텍스트의 활용은 이러한 불확실성 극복의 중요한 열쇠가 된다. 예를 들어, 세미나실에는 빔 프로젝터가 존재하고, 화장실에는 세면 등의 활동을 위한 세면대가 존재한다. 이러한 고수준의 컨텍스트들은 불확실성에 강인한 영상 이해를 도와준다. 또한 고수준의 컨텍스트의 모델링에는 베이지안 네트워크가 좋은 성능을 보이는데 이것은 베이지안 네트워크가 다방향 추론과 불확실성에 강인한 특징을 가지고 있는 확률 모델이기 때문이다[1].

로봇의 이동에 따라 얻는 연속적인 영상에서 장소를 인식하기 위해서는 시간에 따른 각 영상들 사이의 관계가 매우 중요하다. 따라서 강인한 인식을 위해서는 이러한 시간의 흐름을 고려해야 한다. 동적 베이지안 네크워크(DBN)는 베이지안 네트워크의 장점을 가지며, 이전 상태정보를 이용하여 시간의 흐름을 추론이 활용할 수 있는 확률 모델이다[2]. 따라서 이러한 동적 베이지안 네트워크를 이용하면 장소 간 이동의 정확한 인식이 가능하다. 본 논문에서는 영상 전처리를 통하여 추출한 전역 특징값과, SIFT를 이용하여 추출한 물체 존재정보를 고수준의 컨텍스트가 모델링된 동적 베이지안 네트워크를 이용하여 인식하는 기법을 제안한다. 실제 대학 실내 환경에서의 실험을 통하여 제안하는 방법의 성능을 평가하였다.

2. 관련 연구

확률모델을 이용한 영상인식 기법은 비전 기반 영상 이해에서 많이 활용되어 왔다. Torralba 등은 영상에서 추출한 전역 특징 벡터들을 은닉 마르코프 모델을 이용하여 인식하고 이것을 컨텍스트 정보로 이용하여 물체가 존재할 가능성이 높은 영역을 구분하였다[3]. 이 연구에서 제안한 저수준 특징추출은 장소인식에서 좋은 성능을 내며 HMM, BN 등을 이용한 추론에 이용 가능하지만 특징추출 결과에 오류가 좌우되는 단점을 가지고 있다. Marengoni 등은 항공 영상에서 건물을 인식하기 위한 영상 추론 시스템인 Ascender I 을 제안했다[4]. 이 연구에서는 계층적 베이지안 네트워크와 유틸리티 이론을 이용하여 영상 처리기를 선택하며 목표 물체가 존재할 영역을 구분하고 BN추론에 필요한 계산량을 줄였다.

DBN을 이용한 영상 이해도 많은 연구들이 진행되어 왔다. Yongmian Zhang등은 얼굴 영상을 분석하여 저수준의 특징 정보들을 뽑아내고 이를 동적 베이지안 네트워크로 모델링하여 얼굴 표정에서 감정을 인식하였다[5]. 이 연구에서는 시간적 흐름에 따른 감정의 변화를 DBN으로 모델링하여 좋은 결과를 보였다. Ying Luo 등은 스포츠 영상을 DBN을 이용하여 어떠한 동작이 이루어지고 있는지 분석하였다[6]. 이 연구에서는 사람의 몸 각 부분을 베이지안 노드로 모델링하고 시간의 흐름에 따른 이동을 표현하였다. 이와 같이 DBN을 이용한 영상인식기법은 시간적 정보가 중요한 환경에서 좋은 성능을 보이고 있다.

3. 컨텍스트 기반 장소 및 물체 인식

이 장에서는 DBN을 사용한 장소 및 물체 인식을 설명한다.

먼저 영상에서의 전역 특징 벡터 추출을 설명하고 SIFT를 이용한 물체 존재정보 추출을 설명한다. 다음으로 DBN을 이용한 장소 및 물체인식에 대하여 알아본다.

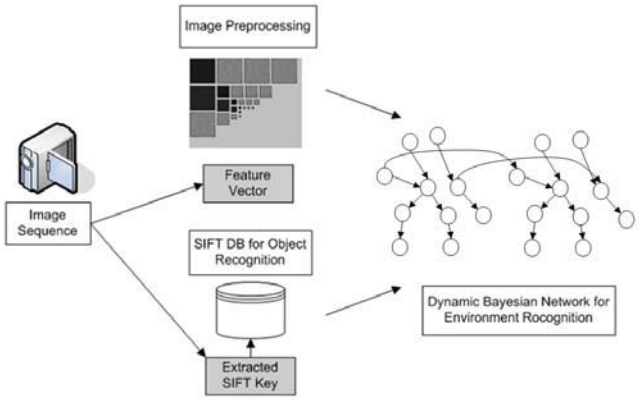

그림 1. 제안하는 방법의 전체적인 구조

3.1 컨텍스트 기반 전역 특징 추출

영상을 이해를 위한 특징을 추출하기 위해서는 영상의 지역적 텍스쳐 성분과 전역적 특징인 영상 구성요소간의 공간적 관계를 모두 검토하는 함수를 이용하여야 한다.[3]. 지역적 특징을 유지하는 텍스쳐 특징을 계산하기 위하여 본 논문에서는 6 방향과 4 스케일의 steerable pyramid를 사용하였다. 시간 t에서의 영상에서 추출되는 벡터는 다음과 같다.

$$v_t^L(x) = v_t, k(x)_{k=1,...,N}, where\ N=24$$

또한 공간적 정보를 유지하는 전역 특징 정보를 갖는 특징의 추출을 위하여 영상의 넓은 공간 영역에서 평균된 등급별 중간값을 사용한다. 특징 벡터를 추출하기 위하여 영상의 해상도를 4x4 pixel로 낮추고 PCA를 이용하여 특징값의 차원을 줄인다. 앞서 설명한 영상처리를 통해 최종적으로 384개의 특징벡터 m_t를 추출한다 (4 x 4 x 24).

$$m_t(x) = \sum_{x'} |v_t^L(x')| w(x'-x),$$
$$where\ w(x)\ is\ the\ averaging\ window.$$

실제 환경에서의 실험을 위하여 USB카메라와 노트북 PC를 입력 영상의 수집에 사용하였다. 사용자가 각 장소를 방문할 때, 사람의 시점 높이에 장착된 카메라를 통하여 320x240 해상도의 영상을 초당 4장씩 캡쳐하여 노트북 PC에 저장한다.

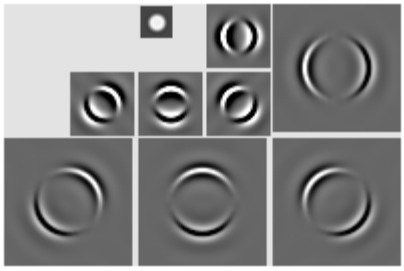

그림 2 steerable pyramid를 이용한 영상 분해의 예

3.2 SIFT를 이용한 물체 존재정보 추출

본 논문에서는 고수준의 물체 존재정보를 추출하기 위하여 크기 불변 특징 변환기법(SIFT)을 이용하였다. 영상정보는 불확실하기 때문에 크기나 카메라의 각도 변화에 강인한 기법이 필요하다. 여러 복잡한 전제하에서 유일하게 가능한 scale-space 커널은 가우시안 함수이다[7]. 입력 영상을 $I(x,y)$라고 할 때 영상의 scale-space $L(x,y,\sigma)$는 다양한 크기의 가우시안 함수들의 컨볼루션 연산을 통하여 다음과 같이 정의된다.

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y)$$

여기서 $*$는 x와 y의 컨볼루션 연산이며 가우시안 함수 G는 다음과 같다.

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

scale-space에서 안정적 키 포인트들을 효과적으로 찾기 위하여 여러 가우시안 함수의 차이로부터의 scale-space extrema를 이용한다. scale-space extrema는 영상 $D(x,y,\sigma)$을 이용하여 계산하며 영상 $D(x,y,\sigma)$는 다음과 같이 구할 수 있다[7].

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y)$$
$$= L(x,y,k\sigma) - L(x,y,\sigma)$$

여기서 k는 상수 곱셈 계수이다.

추출된 각각의 키포인트들을 각각의 영상에서 검토하여 알고리즘의 수행 결과가 임계값을 넘게 되면 SIFT 추출기법은 물체의 존재가 확인되었다고 판단한다. 본 논문에서는 물체 학습 영상들로부터 물체들의 SIFT 키포인트들을 추출하여 XML 데이터베이스에 저장했다. SIFT 학습에 사용한 물체들의 학습 영상은 실제 실험에서 학습에 사용한 이동 영상에서 전문가가 영상처리를 통하여 추출하였다.

3.3 동적 베이지안 네트워크를 이용한 환경 인식

동적 베이지안 네트워크(DBN)는 일련의 확률변수들의 조건부 확률 분포를 표현하는 확률 모델로서 특히 이전 상태정보를 이용하여 시간적인 관계를 표현할 수 있다[2]. DBN은 보통 시간 흐름에 불변인 파라미터들과 전이 종속에 따른 마르코프 특성을 전제로 한다. 이것은 현재 상태의 확률변수는 이전 N단계 혹은 $N-1$단계까지의 영향만을 받고 그 이전의 시간 단계에서는 영향을 받지 않음을 의미한다[2]. 따라서 추론을 위해서는 DBN의 구조에서 N개의 시간 단계만큼 unrolling된 새로운 네트워크 B_1에 의하여 각 확률 변수들의 조건부 확률 분포를 계산해야 한다.

$$p(v_1,...,v_N) = \prod_{n=1}^{N} \prod_{i=1}^{m} p(v_n^i | \pi(v_n^i))$$

본 논문에서 사용한 DBN은 전문가가 설계하였으며 낮은 인과관계를 가지는 노드들은 계산의 복잡성을 줄이기 위하여 연결하지 않았다. 실험에서 사용한 DBN의 마르코프 특성은 바로 전 단계의 결과만을 고려하는 1로 설정하였다.

베이지안 네트워크에 증거값을 설정하기 위해서는 증거값의 이산화가 필요하다. 그러나 영상 전처리를 통하여 추출한 전역 특징 벡터는 확률 증거값이므로 본 논문에서는 가상 증거 기술(virtual evidence technique)을 정의하여 확률 증거값을 DBN에 적용해 주었다. 또한 카메라로 들어오는 모든 영상에 베이지안 추론을 수행하는 것은 너무 많은 계산이 필요하므로 실제 실험에서는 8장의 영상에서 추출한 전역 특징 벡터의 평균값을 증거값으로 활용한다. SIFT를 통하여 추출하는 물체정보 또한 8장에서 추출한 정보를 모아서 DBN에 적용해 주었다.

4. 실험 및 결과

복잡한 실제 환경에서의 실험을 위하여 실험 데이터는 대학

실내 연구 환경에서 USB 카메라를 이용하여 수집하였다. 사용자는 카메라와 함께 다양한 장소를 랜덤하게 방문하며 총 10번의 시퀀스를 수집하였다. 수집된 시퀀스 중에서 5개는 학습, 5개는 실험에 사용되었고 한 시퀀스 당 약 900장의 영상을 수집하였다. 표 1은 실제 실험에서 사용한 장소와 물체들이다.

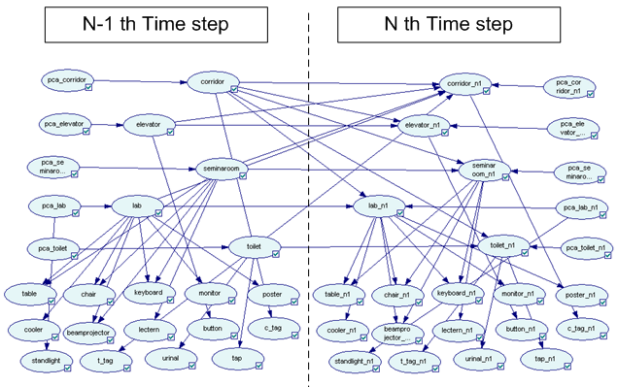

그림 3. 실험에서 사용한 DBN의 구조

표 1. 실험에 사용한 장소와 물체

	종 류
장소(5)	세미나실, 연구실, 화장실, 엘리베이터, 복도
물체(14)	테이블, 의자, 키보드, 모니터, 포스터, 스탠드, 빔 프로젝터, 강의대, 엘리베이터버튼, 변기, 복도태그, 선풍기, 화장실 태그, 수도꼭지

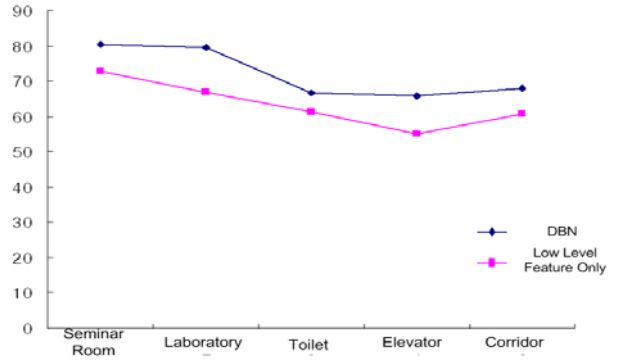

그림 4. DBN을 이용한 장소인식 결과

그림 4는 장소 이동에 따른 실험 결과를 보여준다. 이 결과는 실험에 사용한 5개의 시퀀스 전체의 인식 결과를 합산한 것이다. 전역 특징 벡터만을 사용한 인식에 비하여 제안하는 방법이 좋은 성능을 냈음을 확인할 수 있다. 그러나 두 장소 간의 이동에서 확률값의 변화가 느려 몇몇 잘못된 인식이 수행되었다. 이는 DBN 추론의 계산량의 감소 및 성능 향상을 위하여 설정한 윈도우 크기가 너무 컸기 때문이다.

그림 5, 6은 제안하는 방법의 물체 인식 성능을 보여준다. 그림 6을 통하여 제안하는 방법이 텍스쳐 성분이 적어 SIFT가 잘 인식하지 못하는 여러 물체들의 인식 성능을 높였음을 확인할 수 있다. 그러나 몇몇 물체에서 잘못된 인식 오류가 증가하는 단점을 확인할 수 있었다.

5. 결론 및 향후 연구

실제 대학 실내 환경에서의 실험을 통하여 동적 베이지안 네트워크를 이용한 장소 및 물체 인식이 좋은 결과를 내는 것을 확인할 수 있었다. 전역 특징 정보가 비슷하게 관측되어 장소를 판단하기 어려운 경우라도 제안하는 방법은 고수준의 장소-물체 컨텍스트와 이전 상태정보를 이용하여 올바르게 인식함을 확인할 수 있었다. 그러나 이전 상태 정보를 추론에 활용하는 DBN은 일반적인 BN에 비하여 계산량이 많이 증가한다는 단점이 있다. 향후 좀 더 많은 장소와 물체를 이용한 확장 연구가 필요하며 실제 로봇으로의 적용도 필요하다. 또한 DBN의 성능 개선을 위한 저수준 특징정보의 최적 윈도우 크기에 대한 연구와 계산량의 감소를 위한 다층 접근방법에 대한 연구를 진행하고자 한다.

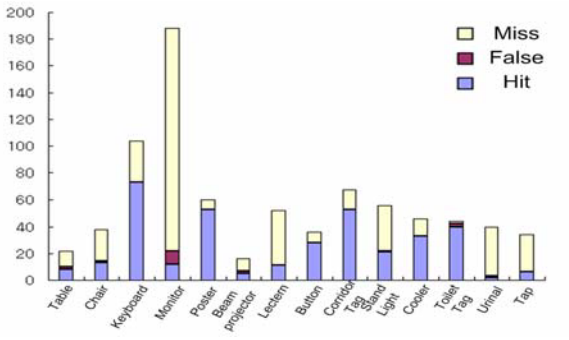

그림 5. SIFT를 이용한 물체인식 결과

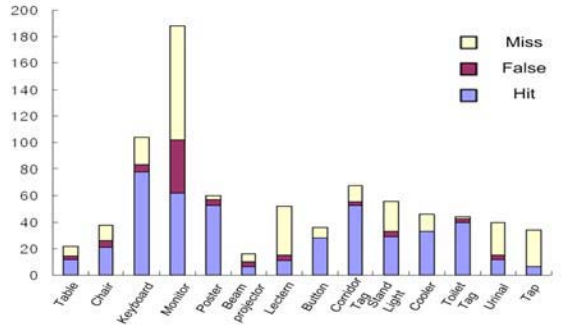

그림 6. DBN 추론을 이용한 물체인식 결과

References

[1] P. Korpipaa, M. Koskinen, J. Peltola, S. Mäkelä, and T. Seppänen "Bayesian approach to sensor-based context awareness," *Personal and Ubiquitous Computing Archive*, vol. 7, no. 4, pp. 113-124, 2003.

[2] K. Murphy, "Dynamic Bayesian networks: Representation, Inference and Learning," PhD thesis, University of California Berkley, 2002.

[3] A. Torralba, K. P. Mutphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *IEEE Int. Conf. Computer Vision*, vol. 1, no. 1, pp. 273, 2003.

[4] . M. Marengoni, A. Hanson, S. Zilberstein and E. Riseman, "Decision making and uncertainty management in a 3D reconstruction system," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 852-858, 2003.

[5] Yongmian Zhang and Qiang Ji, "Active and dynamic information fusion for facial expression understanding from Image Sequences", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 699-714 , 2005.

[6] Ying Luo, Tzong-Der Wu, and Jenq-Neng Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks," *Computer Vision and Image Understanding* vol 92, no. 2, pp. 196-216, 2003

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.