멀티카메라 환경에서의 베이지안 네트워크 기반 이벤트 인식

임수정○ 민준기 박한샘 조성배
연세대학교 컴퓨터과학과
{soojung○, loomlike, sammy}@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Bayesian Network based Event Recognition in Multi-Camera Environment

Soojung Lim○, Junki Min, Han-Saem Park and Sung-Bae Cho
Department of Computer Science, Yonsei University

요 약

기존의 멀티 카메라 시스템은 넓은 영역을 커버하거나 이동 중인 물체를 트래킹 하기 위한 목적으로 주로 사용되어 왔다. 하지만 이러한 시스템은 하나의 카메라가 커버하는 영상이 가려지면 정보를 잃게 되는 단점이 있다. 멀티 카메라 시스템은 하나의 영역을 여러 카메라가 커버하도록 하여 이런 단점을 극복할 수 있다. 또한 다양한 시점의 카메라에서 수집되는 영상의 경우, 영상에 따라 담고 있는 정보가 다르므로 여러 카메라의 입력 정보를 함께 활용하여 보다 많은 정보를 얻을 수도 있다. 본 논문은 이런 장점을 활용하여 멀티 카메라 환경에서의 이벤트 인식 문제를 다룬다. 이를 위해 사무실 환경에 8대의 카메라를 설치하였으며, 시나리오에 따라 영상을 수집하였다. 수집된 영상은 전문가에 의해 어노테이션 된 후 인식 모델의 학습에 사용되며, 학습된 베이지안 네트워크 모델의 구조와 파라미터를 도메인 지식에 기반해서 수정하여 최종 이벤트 인식 모델을 설계하였다. 실험 결과 제안하는 이벤트 인식 모델의 인식률은 평균 87.0%로 Naive Bayes보다 우수한 성능을 보임을 확인하였다.

1. 서 론

최근 카메라의 보급화와 압축, 저장, 디지털화 등 멀티미디어 처리기술의 발전으로 인해 누구나 쉽고 저렴하게 동영상 데이터들을 이용할 수 있게 되었다[1]. 이는 공공건물에서의 CCTV를 비롯해 개인용 휴대 단말기에 이르기까지 많은 동영상들을 수집 가능하게 하였는데, 이러한 동영상 데이터들은 텍스트문서나 음성, 그림파일 등의 다른 데이터 보다 훨씬 구체적이고 사실적이며 정확한 정보를 포함하고 있어 활용가치가 높다.

방송의 대상이 되는 무대나 대형 매장, 사무실, 스포츠 경기에서는 넓은 영역과 다양한 각도를 커버하기 위해 많은 수의 카메라가 사용 되고 있다. 동시에 다양한 카메라에서 수집되는 영상의 경우 영상에 따라 담고 있는 정보의 질과 양이 다를 뿐 아니라, 카메라의 시각의 겹치는 영역이 발생하기 때문에, 사용자가 원하는 정보에 가장 근접하는 정확하고 자세한 영상을 분석하는 작업을 필요로 한다.

본 논문에서는 사무실 환경에서 수집된 영상 내의 이벤트 인식에 초점을 맞추어, 여러 대의 카메라에서 수집된 정보를 바탕으로 사람·사물 등의 개체와 이들 사이에서 발생하는 이벤트를 인식하고자 하였다. 이를 위해 도메인 지식을 바탕으로 베이지안 네트워크 모델을 설계하고 테스트하였다. 시나리오를 설계하고 실험을 수행하기 위해, 사무실 환경에서 발생 가능한 여러 이벤트들을 사전에 정의하였으며, 카메라는 동일한 이벤트를 여러 방향에서 비추도록 설치하였다. 이러한 멀티카메라를 통해 수집한 영상은 전문가에 의해 수동으로 어노테이션하여 모델의 학습 및 검증에 활용하였다.

2. 관련 연구

2.1. 멀티카메라 시스템

기존 멀티카메라 연구에서 멀티카메라 환경은 주로 넓은 영역을 커버하거나, 이동중인 사람, 물체를 추적하기 위해 사용되었다. J. Black 등은 실외환경에서의 움직이는 물체를 추적하고 검출하기 위한 멀티카메라 시스템을 개발하였고[2], 최근에는 실내 환경에서의 멀티카메라 시스템에 대한 연구를 진행 중이다. Y. Sumi 등은 학회장에 설치되어진 센서와 멀티카메라를 이용해 피험자들의 행동을 분석하였고[3], G.C. de Silva 등은 유비쿼터스 환경 내에서 가정 내 구성원들의 정보를 요약하거나 검색하기 위해 멀티카메라를 사용하였다[4].

멀티 카메라 시스템이 갖는 또 하나의 장점은 한 명의 피험자 또는 하나의 이벤트에 대한 다양한 시각을 가진 영상을 얻을 수 있다는 것이다. 이러한 시스템에서는 카메라가 커버할 수 있는 영역을 벗어나는 사각지대나, 사람과 물체의 방향에 따라 사용자가 원하는 영상이 가려져서 잘 보이지 않는 경우 등이 발생하지 않아, 대부분의 경우 원하는 영상정보를 얻을 수 있다. 본 논문에서는 기존의 연구에서 다루어 지지 않은 이러한 가능성에 초점을 맞추어 연구를 진행하였다.

2.2. 베이지안 네트워크

베이지안 네트워크는 최근 복잡한 도메인 영역에서 불확실성을 다루는 확률 모델로 널리 사용되고 있다[5]. 이는 모델링 과정에서 전문가의 지식을 사용할 수 있어, 사람의 움직임과 이벤트 등에 대한 모델을 제시하는데 큰 도움을 주기 때문에 영상처리 분야에서 널리

활용되고 있다[6]. 또한 완벽하지 않은 데이터로도 추론이 가능해 필요한 정보에 대한 확률값을 얻을 때 유용하게 사용된다[7]. 베이지안 네트워크는 방향성 비순환 그래프로써, 베이즈 규칙을 사용하여 도메인 변수들 간의 확률적 의존성을 나타내는 모델이다[8]. 노드간의 관계는 원인이 되는 노드와, 그 영향을 받는 결과 노드로 구분할 수 있는데, 결과 노드는 자신에 대한 초기 확률값 또는 원인노드와의 의존성을 고려한 조건부 확률 테이블을 가진다.

S. Gong등은 베이지안 네트워크를 통해 자동으로 인간의 행동 패턴을 인식하는 모델을 만들었으며[9], Y. Luo 등은 스포츠 비디오 내에서 사람의 머리, 손, 발 등의 좌표로 전체의 움직임을 분석하고 인식하는 모델을 설계하기 위해 동적 베이지안 네트워크(Dynamic Bayesian Network)를 사용하였다[10]. S. Hongeng 등은 한명 또는 여러 명의 사람을 인식하기 위해 베이지안 네트워크와 확률론에 기반한 유한상태 오토마타(Finite state Automata)를 사용하였다[11].

본 연구에서는 베이지안 네트워크를 이용한 모델링 과정에서 도메인 지식을 사용하여 이벤트 노드와 증거 노드 간 계층을 설정해 정확도를 높이는 등의 차별성을 두었다.

3. 베이지안 네트워크 기반 이벤트 인식 모델링

3.1. 실험환경 설정

사무실 내에서의 실제 영상 데이터 수집을 위해, 연구실 내에 4m×4m 영역을 목적영역으로 설정하고 8대의 카메라를 설치하였다. 그림 1은 실험환경 내에서의 카메라의 위치와 카메라가 비추고 있는 영역을 보여주며, 그림 2는 8대의 카메라를 통해 실제로 수집한 영상을 보여준다.

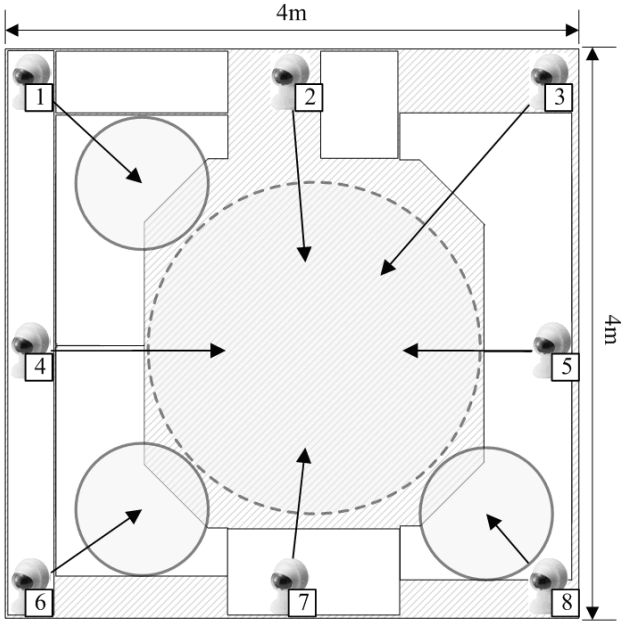

그림 1. 카메라의 위치와 시각 영역

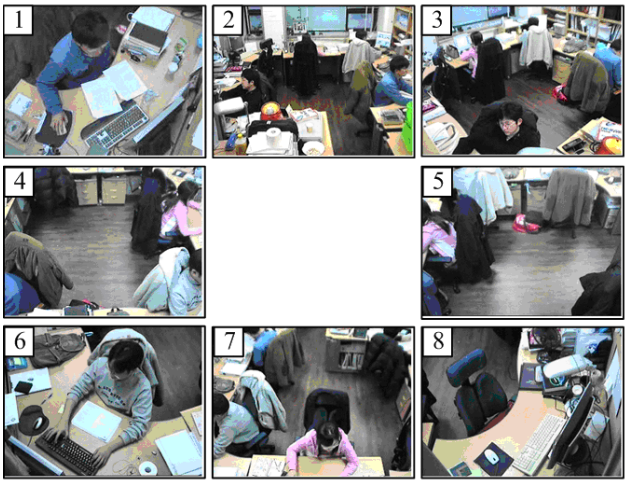

그림 2. 멀티 카메라를 통해 수집한 영상

실험에 사용되어진 카메라는 소니 네트워크 카메라(SNC-P5)로써, 영상은 320×240의 해상도에 15fps의 MPEG 동영상 포맷으로 저장하였다.

3.2. 이벤트 정의

이벤트 인식 모델 설계를 위한 데이터를 수집하기 전에, 사무실 환경 내에서 발생할 수 있는 8가지 이벤트를 사전에 정의하였다. 하나의 이벤트가 발생하기 위해서는 피험자와 피험자가 사용하고 있는 물체와의 관계, 시선방향, 이벤트가 발생하고 있는 위치 등을 알아야 하므로, 이벤트를 정의하는 과정에서 피험자 수, 사용하고 있는 물체, 위치, 시선방향, 자세 등을 고려하였다. 물체의 위치는 실험 영역을 4×4 격자로 나누고, 위에서부터 차례로 1부터 16번까지의 번호를 이용해 구분한다. 정의된 이벤트는 아래와 같다.

- Calling(A, *phone*) if location(A, 1|2|5)
- Cleaning(A, *vacuum machine*)
- Conversation(A, B)
- Meeting(All, *note*)
- Presentation(All, *computer*) if location(A, 1|4|16)
- Work(A, *computer*) if location(A, 1|4|16)
- Study(A, *note*) if location(A, 1|4|16)
- Sleeping(A) if pose(A, rest)

3.3. 이벤트 어노테이션

이벤트의 인식을 위해서는 그 단서가 되는 사람이나 물체의 인식이 선행되어야 하는데, 이를 위해서는 SIFT 등의 알고리즘을 이용하기도 하지만[12], 어노테이션 프로그램을 이용하여 영상 내에서 보이는 장면, 물체 등에 대한 정보를 사람이 직접 어노테이션하기도 한다[13].

본 논문에서도 앞의 이벤트 정의에 따라 이벤트 발생 시의 피험자 ID, 물체, 위치, 시선방향, 자세, 이벤트에 대해 전문가가 직접 어노테이션을 수행하였다. 어노테이션 된 데이터는 이벤트 인식의 단서로 쓰임과 동시에 BN 설계에 앞선 참고용 모델의 학습을 위해서 사용된다.

그림 3은 실제 어노테이션에 사용된 툴을 보여준다.

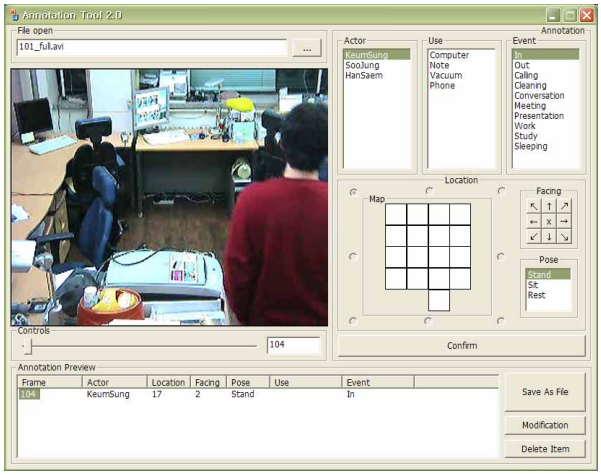

그림 3. 어노테이션 툴

3.4. 이벤트 인식 모델

이벤트 인식 모델의 설계를 위해, 8대의 카메라로부터 얻어진 데이터들을 어노테이션 정보를 바탕으로 통합하여 학습 데이터를 생성하였다. 정확한 이벤트 모델을 생성하기 위해서는 이벤트의 대한 발생 가능한 모든 경우의 학습데이터가 있어야 한다. 실제로 본 논문에서는 사무실 환경 내의 일상적 행동에 대한 시나리오에 기반하여 데이터를 수집하였기 때문에, 시나리오 상에는 없지만 실제로는 발생 가능한 이벤트에 한해 데이터를 추가하였다. 또한 데이터 수에 따라 모델의 학습 결과가 영향을 받을 수 있으므로, 일반적인 모델 생성을 위해 이벤트별 데이터 수를 50개로 맞추었다.

이렇게 구성한 학습데이터를 이용하여 구조 및 파라미터 학습을 하였으며, 이때 도메인 지식을 사용하여 이벤트 노드와 증거노드의 계층을 설정하였다. 학습의 결과로 얻어진 모델의 네트워크 구조는 노드별 조건부 확률값을 분석하여 수정하였다. 물체에 관련한 확률 증거 값을 넣어줄 경우, 물체 노드간의 연결은 의미가 없어지므로 이를 제거하였다. 이와 같은 과정을 거쳐 수정된 이벤트 인식 모델을 바탕으로 이벤트가 확실히 일어나는 경우에는 파라미터 값을 0.999999로, 일어나지 않는 경우에는 0.000001로 변경하였다. 완성된 이벤트 인식 모델의 구조는 그림 4와 같다.

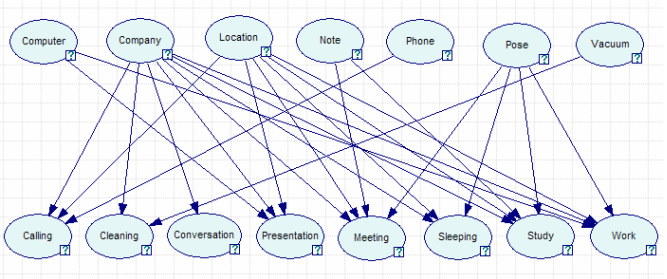

그림 4. 이벤트 인식을 위한 베이지안 네트워크 모델

멀티카메라로부터 수집된 여러 개의 영상을 하나의 영상으로 통합하면, 같은 시간대에 두 명 이상의 피험자가 각각 다른 이벤트를 진행하는 경우가 있어 하나의 모델 구조를 생성 후 이를 각각 피험자에게 독립적으로 적용하였다.

4. 실험 결과

4.1. 데이터 수집

데이터 수집을 통해 얻어진 영상은 사무실 내의 일상생활에서 발생하는 모든 이벤트들에 대해 3명의 피험자에 대한 각각의 시나리오를 설계한 후, 시나리오에 따라 촬영 되었다. 시나리오의 총 길이는 9분이며, 오전 9시부터 6시까지의 업무시간을 기준으로 1시간을 1분으로 가정하였다. 그림 5은 실제 데이터 수집에 사용된 시나리오의 예를 보여준다.

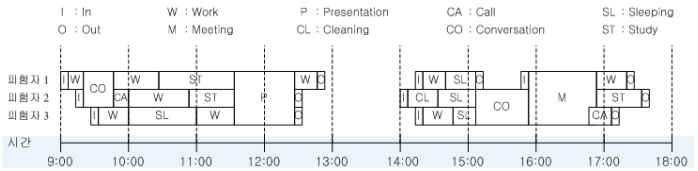

그림 5. 설계된 시나리오의 예

추가 데이터의 수집 역시 다른 시나리오를 설계하여 위와 같은 방법으로 이루어졌다.

4.2 이벤트 인식 실험

제안하는 이벤트 인식 모델의 성능을 검증하기 위해 피험자별 이벤트 인식 실험을 하였다. 인식 실험을 위해 어노테이션 된 데이터에서 이벤트를 제외한 피험자, 물체의 정보를 증거 값으로 넣어주고 이벤트 인식 모델을 통해 이벤트를 추론하였다. 모델은 각 데이터에 대해 8개의 이벤트 중 가장 큰 확률 값을 보이는 이벤트를 인식의 결과로 돌려준다. 이 결과 값과 어노테이션 툴을 사용해 레이블링 된 실제 이벤트를 비교하여 인식률을 계산하였다.

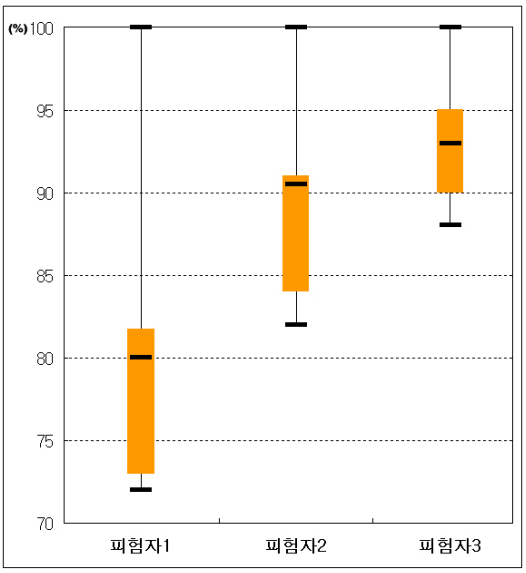

그림 6. 이벤트 인식 실험의 정확도

그림 6은 피험자별 이벤트 인식실험을, 제안하는 모델을 이용하여 10회 반복한 결과를 box plot의 형태로 보여준다. 실험 결과 가장 낮은 인식률을 보인 Presentation은 이벤트 정의를 만족시키지 못하는 영상, 즉 컴퓨터의 사용여부가 불분명할 때 여러 명의 피험자들이 컴퓨터나 다른 물체의 사용 없이 이야기를 하고 있는 Conversation(인식률 67%)라고 인식하는 경우가 대부분이었다.

그림 7은 Naive Bayes와의 성능 비교 평가 결과를 보여주며, 제안하는 이벤트 인식 모델이 상대적으로 더 우수한 성능을 보이는 것을 확인할 수 있다. Naive Bayes는 증거 노드와 이벤트 노드를 모두 연결한 모델이다.

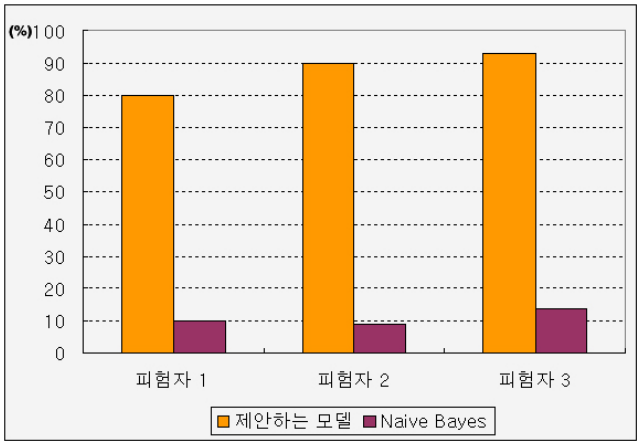

그림 7. Naive Bayes와의 성능 비교 평가

5. 결론 및 향후 계획

본 논문에서는 사무실 환경에 설치된 멀티카메라를 통해 얻어진 영상으로부터 이벤트를 인식하기 위해 베이지안 네트워크 기반 이벤트 인식 모델을 제안하였다. 8대의 카메라로부터 얻어진 영상 정보들을 각각 전문가에 의해 어노테이션 된 후 통합되어 학습데이터로 생성하였고, 학습되어진 베이지안 네트워크 모델의 구조와 파라미터를 도메인 지식에 기반해서 수정하여 최종 이벤트 인식 모델을 설계하였다. 제안한 이벤트 인식 모델의 이벤트 인식률은 86.0%로 Naive Bayes와 비교해 우수한 성능을 보여주었다.

향후에는 이벤트 인식에서 나아가, 인식 결과를 바탕으로 상위 레벨에서의 영상검색에 적용할 수 있을 것이다. 베이지안 네트워크 모델의 결과인 이벤트 별 인식 확률을 활용하면, 효과적인 검색이 가능할 것이다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었음. IITA-2006-(C1090-0603-0046)

참고문헌

[1] M. Petkovi and W. Jonker, "Integrated use of different content derivation techniques within a multimedia database management system," *Journal of Visual Communication and Image Representation*, vol. 15, pp. 303-329, 2004.

[2] J. Black and T. Ellis, "Multi camera image tracking," *Image and Vision Computing*, vol. 24, pp. 1256-1267, 2006.

[3] Y. Sumi, S. Ito, T. Matsuguchi, S. Fels and K. Mase, "Collaborative capturing and interpretation of interactions," *Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pp. 1-7, 2004.

[4] G. C. de Silva, T. Yamasaki and K. Aizawa, "Evaluation of video summarization for a large number of cameras in ubiquitous home," *Proceedings of the 13th ACM International Conference on Multimedia*, pp. 820-828, 2005.

[5] S.-M. Lee and P. A. Abbott, "Bayesian networks for knowledge discovery in large datasets: Basics for nurse researchers," *Journal of Biomedical Informatics*, vol. 36, pp. 389-399, 2003.

[6] 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, pp.852-858, 2003.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, 1988

[8] E. Charniak, "Bayesian networks without tears," *AI Magazine*, vol.12, pp.50-63, 1991

[9] S. Gong, J. Ng, J. Sherrah, "On the semantics of visual behaviour, structured events and trajectories of human action," *Image and Vision Computing*, vol. 20, pp. 873-878, 2002.

[10] Y. Luo, T.-D. Wu and J.-N. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks," *Computer Vision and Image Understanding*, vol. 92, pp. 196-216, 2003.

[11] S. Hongeng, R. Nevatia and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol. 96, pp. 129-162, 2004.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[13] A. Amir, S. Basu, G. Iyengar, C. Lin, M. Naphade, J. Smith, S. Srinivasan and B. Tseng, "A multi-modal system for the retrieval of semantic video events," *Computer Vision and Image Understanding*, vol. 96, pp. 216-236, 2004.