연속적인 손 제스처의 실시간 인식을 위한 계층적 베이지안 네트워크

(A Hierarchical Bayesian Network for Real-Time Continuous Hand Gesture Recognition)

허 승 주 † 이 성 환 ††
(Sung-Ju Huh) (Seong-Whan Lee)

요 약 본 논문은 컴퓨터 마우스를 제어하기 위한 실시간 손 제스처 인식 방법을 제안한다. 다양한 제스처를 표현하기 위해, 손 제스처를 연속적인 손 모양의 시퀀스로 정의하고, 이러한 손 제스처를 인식하기 위한 계층적 베이지안 네트워크를 디자인한다. 제안하는 방법은 손 포스처와 제스처 인식을 위한 계층적 구조를 가지며, 이는 특징 추출과정에서 발생하는 잡음에 강인하다는 장점을 가진다. 제안하는 방법의 유용성을 증명하기 위해, 제스처 기반 가상 마우스 인터페이스를 개발하였다. 실험에서 제안한 방법은 단순한 배경에서는 94.8%, 복잡한 배경에서는 88.1%의 인식률을 보였으며, HMM 기반의 기존 방법보다 우수한 성능을 보였다.

키워드 : 계층적 베이지안 네트워크, 손 제스처 인식, 휴먼-컴퓨터 인터페이스

Abstract This paper presents a real-time hand gesture recognition approach for controlling a computer. We define hand gestures as continuous hand postures and their movements for easy expression of various gestures and propose a Two-layered Bayesian Network (TBN) to recognize those gestures. The proposed method can compensate an incorrectly recognized hand posture and its location via the preceding and following information. In order to vertify the usefulness of the proposed method, we implemented a Virtual Mouse interface, the gesture-based interface of a physical mouse device. In experiments, the proposed method showed a recognition rate of 94.8% and 88.1% for a simple and cluttered background, respectively. This outperforms the previous HMM-based method, which had results of 92.4% and 83.3%, respectively, under the same conditions.

Key words : Hierarchical Bayesian Network, Hand Gesture Recognition, Human-Computer Interface

1. 서 론

지난 10년 동안, 키보드나 마우스 대신에 컴퓨터를 제어하기 위한 다양한 상호작용 방법들이 소개되고 개발되어왔다. 그 중 제스처 기반의 제어 방식은 매우 편리하고 직관적이기 때문에, 인간-컴퓨터 상호작용을 위해 활발하게 연구가 진행되고 있다[1]. 최근의 연구들은 시계열의 제스처 데이터를 효율적으로 모델링하는데 초점을 두고 있고, 은닉 마르코프 모델(HMM), 동적 베이지안 네트워크(DBN)과 같은 다양한 확률 모델들이 소개되었다. T. Starner와 A. Pentland는 HMM을 이용해 미국 수화를 실시간으로 인식하는 방법을 제안하였다[2]. S. Marcel 등은 손 제스처를 인식하기 위해 입출력 은닉 마르코프 모델(IOHMM)[3]을 제안하였으며, 또한 A. El-Sawah 등은 DBN을 통해 손 제스처를 인식하는

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2009-0060113). 이 연구에 참여한 연구자는 '2단계 BK21사업'의 지원을 받았음

† 비 회 원 : 고려대학교 컴퓨터 · 통신공학부
sjheo@image.korea.ac.kr
†† 종신회원 : 고려대학교 컴퓨터 · 통신공학부 교수
swlee@image.korea.ac.kr
(Corresponding author임)

논문접수 : 2009년 7월 22일
심사완료 : 2009년 9월 1일

프레임워크를 제안하였다[4].

본 논문은 복잡한 환경에서도 강인하게 동작하는 실시간 손 제스처 인식을 위해 2-계층 베이지안 네트워크를 제안한다. 하위 계층에서는 베이지안 네트워크를 이용하여 손의 포스처를 인식하고, 상위 계층에서는 동적 베이지안 네트워크를 이용하여 손 제스처를 인식한다. 이 계층적 모델은 포스처를 위한 특징과 제스처를 위한 특징에 대한 의존성을 분리하기 때문에 확률 추론 과정에서의 계산 복잡도를 줄일 수 있다. 또한, 하위 계층에서 복잡한 환경 등과 같이 조악한 조건 때문에 손 포스처가 올바르게 인식되지 않아도, 상위 계층에서 이전 시간의 정보와 예측되는 정보를 이용하여 이를 보완할 수 있다.

2. 계층적 베이지안 네트워크

제안하는 손 제스처 인식 방법은 동적 베이지안 네트워크[5]와 그 확장에 기반을 둔다. 동적 베이지안 네트워크는 음성, 제스처 등과 같은 다양한 시계열 데이터를 인식하는데 있어 가장 표현력 있는 모델 중 하나이다.

본 논문에서는, 연속적인 손 제스처를 모델링하기 위해 계층적 베이지안 네트워크를 제안한다. 손 제스처는 손 포스처와 그 이동에 대한 시퀀스로 구성되기 때문에, 포스처와 손의 이동 정보는 입력 영상으로부터 추출되는 특징과 손 제스처 사이의 중간 파라미터라고 할 수 있다. 이런 측면에서, 손 제스처 인식 문제는 손의 포스처를 인식하는 하위 단계와 포스처의 시퀀스와 손의 이동 방향을 인식하는 상위 단계로 나누어 모델링할 수 있다. 하위 계층은 시간 특징을 입력으로 하고, 손 포스처의 인식 결과를 출력으로 하는 베이지안 네트워크로 구성된다. 상위 계층은 하위 계층의 결과와 손의 움직임에 대한 방향 벡터를 입력으로 하고, 손 제스처 인식 결과를 출력으로 하는 동적 베이지안 네트워크로 구성된다. 다시 말해, 동적 베이지안 네트워크는 하위 계층의 결과인 손 포스처의 시간적인 변화를 나타낸다. 이 계층적 모델은 포스처와 제스처를 위한 특징들에 대한 의존성을 분리하여 확률 추론 시 계산 복잡도를 줄여주고, 시간적인 제약을 통해 잘못 추출된 시각 특징들에 의한 오류를 보정할 수 있게 한다. 계층적 베이지안 네트워크 설명에 사용되는 기호들의 의미는 표 1에 정의되어 있다.

포스처를 모델링하기 위해, 하위 계층에 특정한 형태의 베이지안 네트워크를 제안한다. 그림 1에서 보는 것처럼, 제안하는 베이지안 네트워크는 손 포스처에 대한 하나의 은닉 노드 X^t와 다섯 손가락에 대한 다섯 개의 은닉 노드들 $\{F^1_t, F^2_t\}$로 구성된다. 이 다섯 은닉 노드들은 다섯 손가락의 관측을 의미하는 관측 노드 $\{O^1_t, O^2_t\}$를 가진다. 검지는 가장 중요하고, 빈번하게 사용되는 손

표 1 계층적 베이지안 네트워크에 사용된 변수 정의

용어	설 명
S^i_t	시간 t 에서 i 번째 손가락에 대한 특징(손가락 길이)
O^1_t	시간 t 에서 검지에 대한 관측 $O^1_t = S^2_t$
O^2_t	시간 t 에서 검지를 제외한 나머지 네 손가락에 대한 관측 $O^2_t = [S^1_t, S^3_t, S^4_t, S^5_t]^T$
G_t	시간 t 에서 손 이동의 방향에 대한 관측
F^1_t	시간 t 에서 검지의 은닉 상태 $F^1_t = F^{index}_t$
F^2_t	시간 t 에서 검지를 제외한 나머지 네 손가락에 대한 은닉 상태 $F^2_t = [F^{thumb}_t, F^{middle}_t, F^{ring}_t, F^{little}_t]^T$
X_t	시간 t 에서 하위 계층 베이지안 네트워크의 은닉 상태
M_t	시간 t 에서 상위 계층 동적 베이지안 네트워크의 은닉 상태

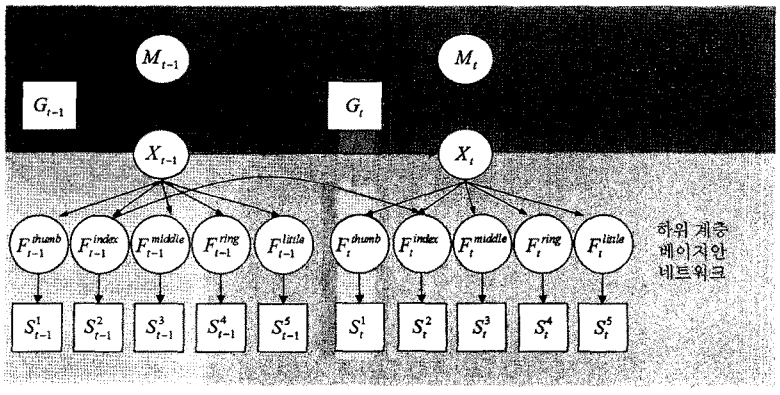


그림 1 계층적 베이지안 네트워크의 도식적 표현

가락이기 때문에, 검지에 대한 은닉 노드는 다른 손가락과는 분리되어 있다. 다른 손가락에 대한 노드들과는 다르게 검지에 대한 은닉 노드는 검지에 대한 상태 변이를 의미하는 이전 시간 t-1로부터의 관계가 이어져 있다.

하위 계층 베이지안 네트워크는 다섯 노드들 $\{O^1_t, O^2_t\}$가 관측되었을 때, 은닉 노드 X_t의 상태를 추론한다. 그림 1에 표현된 베이지안 네트워크의 위상에 따라, 확률 분포 $P(X_t|O^1_{1:t},O^2_{1:t})$는 다음과 같이 정리된다.

$$P(X_t|O^1_{1t},O^2_{1t}) = \sum_{F^1}\sum_{F^2} P(O^1_t|F^1_t)P(O^2_t|F^2_t)P(F^1_t|X_t)P(F^2_t|X_t)P(F^1_t|F^1_{t-1}) \sum_{X_{t-1}} P(X_t|X_{t-1})P(X_{t-1}|O^1_{1t-1},O^2_{1t-1})$$

여기서, 첫번째 항 $P(O^1_t|F^1_t)$는 검지의 관측 데이터에 대한 확률 분포, 두 번째 항 $P(O^2_t|F^2_t)$는 다른 네 손가락들의 관측 데이터에 대한 확률 분포, 세 번째 항 $P(F^1_t|X_t)$는 포스처의 상태 X_t에 대한 검지 F^1_t의 조건부 확률, 네 번째 항 $P(F^1_t|F^1_{t-1})$는 1차 마르코프 과정을 이용한 검지 F^1_t의 조건부 확률, 다섯 번째 항 $P(F^2_t|X_t)$는 포스처의 상태 X_t에 대한 다른 손가락들 F^2_t의 조건부 확률, 여섯 번째 항 $P(X_t|X_{t-1})$은 포스처 상태의 변이 확률 분포, 마지막 항은 $P(X_{t-1}|O^1_{1:t-1}, O^2_{1:t-1})$의 t-1시간에서의 형태로 이를 통해 반복적으로 계산된다.

제스처를 모델링하기 위해, 상위 계층에 동적 베이지안 네트워크를 적용한다. 제안하는 동적 베이지안 네트워크는 그림 1과 같이, 은닉 노드 M_t와 두 개의 관측 노드 X_t, G_t로 구성된다. X_t는 손 포스처의 확률 벡터이자 하위 계층 베이지안 네트워크의 출력이고, G_t는 손의 이동을 움직임 방향 히스토그램을 이용해 얻은 8 방향의 벡터이다. 하위 계층의 베이지안 네트워크와 상위 계층의 동적 베이지안 네트워크를 연결시키는 과정은 다음과 같다. 하위 계층 베이지안 네트워크는 각 포스처에 대한 확률을 출력하고, 이 확률들을 벡터화하여 동적 베이지안 네트워크의 관측으로 사용하기 위한 벡터로 재구성한다. 이 관측 벡터 X_t는 손의 방향에 대한 관측 G_t와 함께 제스처에 대한 은닉 상태 M_t에 대한 확률 분포를 추정한다.

동적 베이지안 네트워크의 확률 $P(M_{t=j}|X_t, G_t)$을 계산하기 위해, 제안하는 동적 베이지안 네트워크에 전방향 알고리즘[5]을 적용하여, 반복적으로 $P(M_{t=j}|X_{1:t}, G_{1:t})$을 계산할 수 있다.

$$P(M_t = j|X_{1:t},G_{1t}) = \frac{1}{c_t}P(M_t = j, X_t, G_t|X_{1t-1},G_{1t-1})$$

여기서,

$$P(M_t = j, X_t, G_t|X_{1t-1},G_{1t-1}) = \left\{\sum_i P(M_t = j|M_{t-1} = i)P(M_{t-1} = i|X_{1t-1},G_{1t-1})\right\}P(X_t,G_t|M_t = j)$$

이고,

$$c_t = P(X_t,G_t|X_{1t-1},G_{1t-1}) = \sum_k P(M_t = k, X_t, G_t|X_{1t-1},G_{1t-1})$$

이다.

3. 실험 결과 및 분석

3.1 실험 환경

학습 및 테스트 데이터는 CMOS 카메라를 이용해 320×240의 24비트 칼라 영상으로 캡처하였다. 데이터 셋은 파란 천을 두른 단순한 배경과 사무실 환경의 복잡한 배경의 2개의 다른 조건에서 촬영된 총 60개의 비디오 클립으로 구성되어 있다. 그림 2는 복잡한 배경에서의 실험 예를 보여주고 있다.

3.2 포스처 및 제스처 인식

제안하는 계층적 베이지안 네트워크의 유용성을 증명하기 위해, 제스처 기반 마우스 인터페이스를 구현하였다. 이를 통해 사용자는 몇 가지 정의된 간단한 손 모양과 제스처를 통해 마우스 장치의 기능을 완벽히 대체할 수 있다. 손 제스처는 간단한 손 모양의 연속적인 변화

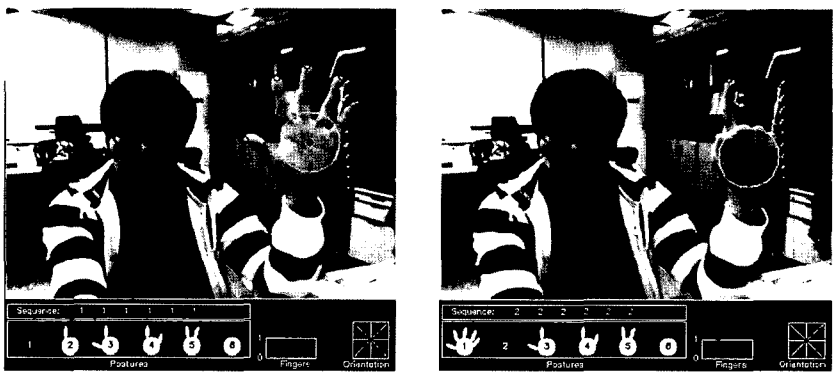


그림 2 사무실 환경의 복잡한 배경에서의 실험

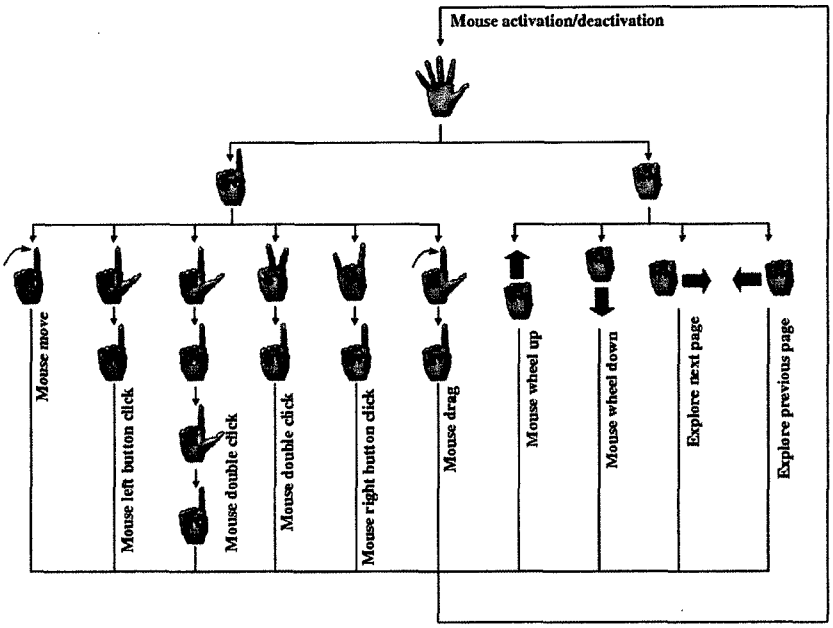


그림 3 마우스 기능 대체를 위한 손 제스처 정의

로서 정의되는데, 손 제스처에 대한 문법은 그림 3에 명시되어 있다. 모든 제스처는 특정 손가락을 쥐었다 펴는 간단한 동작으로 수행하거나 특정 손 모양으로 간단한 궤적을 그림으로써 수행된다.

3.3 인식 결과 및 분석

제안하는 계층적 베이지안 네트워크의 성능 평가를 위해, 유한 상태 머신(FSM)[6]과 은닉 마르코프 모델(HMM)[7]에 기반을 둔 기존의 방법들과 비교평가를 수행하였다. 기존의 방법들은 그들의 논문에 설명된 것과 동일하게 구현되었고, 2가지 배경 조건의 동일한 데이터 셋을 사용해 테스트 하였다. 계층적 베이지안 네트워크는 실험을 통해 결정된 고정된 크기의 슬라이딩 윈도우를 사용해 테스트 하였다.

그림 4는 마우스 드래그 이벤트에 대한 손 제스처를 포함하는 영상 시퀀스에 대한 2계층 인식 과정에 대한 예시를 보여준다. 복잡한 배경 때문에, 포스처 인식을 위한 특징들에 잡음이 포함되어 있음을 그림 4(a)에서 확인 할 수 있다. 그림 4(b)에서, 하위 계층에서 베이지안 네트워크를 통해 포스처를 인식하지만 프레임 15와 24에서와 같이 잘못 인식되는 오류가 발생한다. 이러한 오류를 상위 계층에서 이전 시간과 예측되는 정보 및 손의 움직임을 사용하여 그림 4(c)와 같이 보정할 수 있다. 마지막으로 2절에 설명된 방법에 따라 가장 높은 확률을 가지는 제스처를 인식한다.

제안하는 계층적 베이지안 네트워크의 정확도를 측정하기 위해, 대체 오류, 삽입 오류, 삭제 오류의 3종류의 오류를 계산하였다. 대체 오류는 입력 제스처가 잘못 인식되었을 경우에 발생하고, 삽입 오류는 존재하지 않는 제스처를 인식할 경우에 발생하고, 삭제 오류는 존재하는 제스처를 인식하지 못할 경우 발생한다. 인식률 RR은 아래와 같이 계산된다.

$$RR = \frac{N-S-I-D}{N} \times 100 = \frac{C-I}{N} \times 100$$

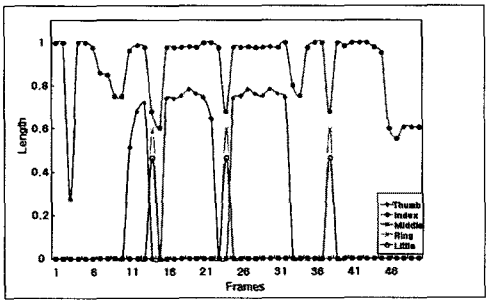
여기서, N은 테스트 데이터의 수, S는 대체 오류의 수, I는 삽입 오류의 수, D는 삭제 오류의 수, C는 $N-(S+D)$로 올바르게 인식된 제스처의 수이다.

기존 방법들과의 비교는 표 2에 보인다. 표 2를 보면, 유한 상태 머신[6]과 은닉 마르코프 모델[7]에 기반을 둔 기존 방법들과 비교해서 복잡한 배경 하에서도 인식률이 많이 떨어지지 않음을 알 수 있다. 이는 계층적 베이지안 네트워크가 하위 계층에서 손 포스처를 정확하게 인식하지 못하더라도 상위 계층에서 이를 보정해 주기 때문이다.

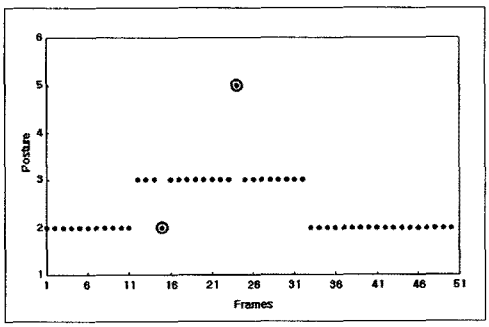
제안한 방법을 이용한 마우스 인터페이스 시스템은 7개의 연속된 프레임을 윈도우로 하여 4프레임씩 겹쳐 이동하면서 연속된 제스처를 인식하였다. 여기서 사용된 윈도우의 크기는 실험을 통해 결정하였다. 마우스 인터페이스 시스템은 펜티엄 Core 2 Duo 2.4GHz CPU와 2GB의 메모리가 장착된 컴퓨터에서 초당 7프레임의 처리 속도를 보였다.

표 2 제안하는 계층적 베이지안 네트워크와 기존 방법들과의 성능 비교

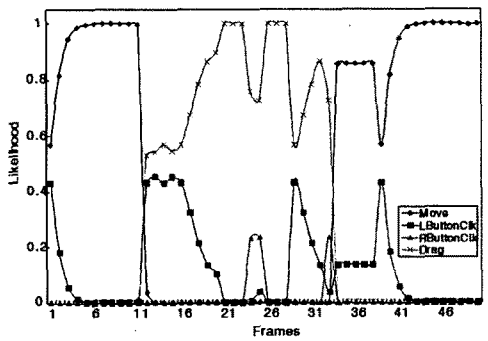
방법	단순한 배경 조건						복잡한 배경 조건					
	N	*C*	오류			*RR* (%)	*N*	*C*	오류			*RR* (%)
			S	*I*	*D*				*S*	*I*	*D*	
FSM	330	315	14	29	1	86.7	330	295	35	34	0	79.0
HMM	330	319	11	14	0	92.4	330	303	25	28	2	83.3
제안 방법	330	321	5	4	0	96.8	330	314	8	8	0	93.9



(a) 추출된 손가락에 대한 특징



(b) 하위 계층 베이지안 네트워크에서의 포스처 인식 결과



(c) 상위 계층 동적 베이지안 네트워크에서의 우도값 변화

그림 4 손 제스처의 계층적인 인식 과정

4. 결 론

본 논문은 계층적 베이지안 네트워크에 기반을 둔 실시간 손 제스처 인식 방법을 제안하였다. 제안한 방법은 계층적인 인식 방법을 통해 시각 특징이 올바르게 추출되기 어려운 복잡한 환경에서도 신뢰할 만한 성능을 보였다. 하위 계층에서 손 포스처를 베이지안 네트워크로 인식하고, 상위 계층에서 연속적인 손 제스처와 그 이동 방향을 동적 베이지안 네트워크를 통해 인식하였다. 이러한 계층적인 인식 방법은 특징간의 의존성을 분리해 계산 복잡도를 줄이고, 하위 계층에서 잘못 인식된 정보들을 상위 계층에서 보정함으로써 보다 나은 성능을 보였다.

참 고 문 헌

[1] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol.37, no.3, pp.311-324, 2007.

[2] T. Starner and A. Pentland, "Real-time American Sign Language Recognition from Video using Hidden Markov Models," MIT Media Lab., MIT, Cambridge, MA, Tech. Rep. TR-375, 1995.

[3] S. Marcel, O. Bernier, J. Viallet, and D. Collobert, "Hand Gesture Recognition using Input-Output Hidden Markov Models," *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, pp.456-461, May 2000.

[4] A. El-Sawah, N. Georganas, and E. Petriu, "A Prototype for 3-D Hand Tracking and Posture Estimation," *IEEE Trans. on Instrumentation and Measurement*, vol.57, no.8, pp.1627-1636, 2008.

[5] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph. D dissertation, University of California, Berkeley, 2002.

[6] M. Yeasin and S. Chaudhuri, "Visual Understanding of Dynamic Hand Gestures," *Pattern Recognition*, vol.33, no.11, pp.1805-1817, 2000.

[7] A. Ramamoorthy, N. Vaswani, S. Chaudhury, and S. Banerjee, "Recognition of Dynamic Hand Gestures," *Pattern Recognition*, vol.36, no.9, pp.2069-2081, 2003.

허 승 주

2007년 건국대학교 소프트웨어학과(학사)
2009년 고려대학교 컴퓨터학과(석사). 관심분야는 컴퓨터 시각, 패턴인식, 영상처리 등

이 성 환

정보과학회논문지 : 소프트웨어 및 응용
제 36 권 제 1 호 참조