동적 베이스망 기반의 양손 제스처 인식
(Dynamic Bayesian Network based Two-Hand Gesture Recognition)

석 흥 일 † 신 봉 기 ††
(Heung-Il Suk) (Bong-Kee Sin)

요 약 손 제스처를 이용한 사람과 컴퓨터간의 상호 작용은 오랜 기간 많은 사람들이 연구해 오고 있으며 커다란 발전을 보이고 있지만, 여전히 만족스러운 결과를 보이지는 못하고 있다. 본 논문에서는 동적 베이스망 프레임워크를 이용한 손 제스처 인식 방법을 제안한다. 유선 글러브를 이용하는 방법들과는 달리, 카메라 기반의 방법에서는 영상 처리와 특징 추출 단계의 결과들이 인식 성능에 큰 영향을 미친다. 제안하는 제스처 모델에서의 추론에 앞서 피부 색상 모델링 및 검출과 움직임 추적을 수행한다. 특징들간의 관계와 새로운 정보들을 쉽게 모델에 반영할 수 있는 동적 베이스망을 이용하여 두 손 제스처와 한 손 제스처 모두를 인식할 수 있는 새로운 모델을 제안한다. 10가지 독립 제스처에 대한 실험에서 최대 99.59%의 높은 인식 성능을 보였다. 제안하는 모델과 관련 방법들은 수화 인식과 같은 다른 문제들에도 적용 가능할 것으로 판단된다.

키워드 : 손 제스처 인식, 손 추적, 동적 베이스망, coupled 은닉 마르코프 모델

Abstract The idea of using hand gestures for human-computer interaction is not new and has been studied intensively during the last decade with a significant amount of qualitative progress that, however, has been short of our expectations. This paper describes a *dynamic Bayesian network* or DBN based approach to both two-hand gestures and one-hand gestures. Unlike wired glove-based approaches, the success of camera-based methods depends greatly on the image processing and feature extraction results. So the proposed method of DBN-based inference is preceded by fail-safe steps of skin extraction and modeling, and motion tracking. Then a new gesture recognition model for a set of both one-hand and two-hand gestures is proposed based on the dynamic Bayesian network framework which makes it easy to represent the relationship among features and incorporate new information to a model. In an experiment with ten isolated gestures, we obtained the recognition rate upwards of 99.59% with cross validation. The proposed model and the related approach are believed to have a strong potential for successful applications to other related problems such as sign languages.

Key words : Hands gesture recognition, hands tracking, dynamic Bayesian network, coupled hidden Markov model

· 이 논문은 2004년도 부경대학교 연구년 교수지원사업에 의하여 연구되었음
· 이 논문은 2007 한국컴퓨터종합학술대회에서 '동적 베이스망 기반의 양손 제스처 인식'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 부경대학교 컴퓨터공학과
daedalos@pknu.ac.kr
†† 정 회 원 : 부경대학교 컴퓨터멀티미디어공학부 교수
bkshin@pknu.ac.kr

논문접수 : 2007년 10월 2일
심사완료 : 2008년 3월 7일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제4호(2008.4)

1. 서 론

사람의 동작은 프로파일, 궤적 등과 같은 움직임에 대한 정보만으로 식별될 수 있음을 Johansson[1]의 실험에서 밝혀진 뒤, 컴퓨터 시각 분야에서도 사람의 행동에 대한 분석 및 이해에 대한 연구[2]가 활발히 진행되고 있으며, 손 제스처 인식에 대한 연구는 Pavlovic 등[3]의 소개 논문에 잘 나타나 있다.

손 제스처는 시간의 흐름에 따른 손의 위치 변화로 표현될 수 있는데 이러한 시간 상의 변화 패턴을 모델링하는데 있어서 동적 프로그래밍(DTW)과 은닉 마르코프 모델(HMM)은 음성 인식, 컴퓨터 시각 등의 여러

분야에서 널리 이용되어 왔다. DTW는 미리 정의된 템플릿과 입력 패턴간의 유사성을 동적 프로그래밍 기법으로 비교하여 인식하는 방법으로 음성 인식에 많이 적용되었다[4]. 이론적으로 간단한 방법이긴 하지만 최적의 템플릿을 찾는 문제와 계산량의 문제를 수반하며, 인식 대상이 많은 경우 템플릿의 개수도 증가하는 단점이 있다. HMM 또한 음성 인식에서 처음 적용되었으며, 컴퓨터 시각 분야에서는 테니스 동작 인식[5], 걸음걸이를 이용한 생체 인식[6], 손 제스처를 이용한 파워 포인트 제어[7] 등 다양한 응용에 적용되어 왔다. Brand 등은 두 손을 이용한 중국 무술 동작 인식을 위해 전형적인 HMM에 구조적 변화를 취한 coupled HMM을 제안하였다[8]. 최근, 확률 모델 중 하나이며, HMM과 Kalman filter의 일발화된 모델인 동적 베이스망(Dynamic Bayesian Network: DBN)에 많은 관심이 집중되고 있다. 동적 베이스망은 베이스망(Bayesian network: BN)[9]의 확장으로 시간 정보가 추가된 모델이다.

Du 등은 두 사람간에 일어날 수 있는 5가지 동작들을 정의하고, 지역적 특성(contour, 모멘트, 높이 등)과 전역적 특성(속도, 방향 등)을 동적 베이스망의 관측값으로 하여 인식을 수행하였으며[10], Park 등은 사람의 신체 각 부분들의 자세 또는 위치 변화를 분석하고, 두 사람간의 상호 작용을 인식하기 위해 동적 베이스망을 적용하였다[11]. Avilés-Arriaga 등은 10 가지의 한 손 동작들에 대해 손 영역의 크기, 중심 좌표, 위치 등의 변화를 특징값으로 하고, 간단한 DBN의 형태인 동적 naïve 베이스 식별기를 인식 모델로 하였고[12], Pavlovic는 HMM과 동적 선형 시스템을 결합한 형태의 동적 베이스망을 이용한 제스처 인식에 대한 연구를 수행하였다[13]. Wilson은 동적 베이스망을 이용하여 여러 분류된 상황들에 적응할 수 있는 제스처 인식 기법 및 모델링을 제안하였다[14]. Oliver 등은 제스처 인식에 있어 HMM과 DBN의 비교 실험에서 HMM에서 서로 독립이라고 가정되었던 확률 변수들간의 관계를 DBN을 이용하여 의존성을 학습할 수 있음을 보였다[15]. 한편, León 등은 15개의 연속한 프레임을 하나의 윈도우로 정의하고, 윈도우 내에서의 손의 위치 변화를 BN에서 하나의 노드로 표현하였으며, 특정 노드의 값이 관측되지 않더라도 “Good-Bye”와 “Move Right”의 동작을 구분할 수 있음을 보였다[16]. Yang 등은 손의 움직임 패턴을 표현하는 궤적 정보를 특징값으로 하여, 연속된 사건들의 시간적 관계를 표현하는 각 계층들간에 있어서 시간적 지연을 적용한 다층 신경망인 TDNN(time-delayed neural network)을 인식 모델로 적용하여 40개의 미국 수화 언어에 대한 실험에서 96.21%의 성능을 보였다[17]. 시청각 음성 인식(Audio-Visual Speech Recognition: AVSR)에 대해 coupled HMM과 factorial HMM 모델을 적용하여 인식을 수행한 Nefina 등은 기존의 화자 독립적인 AVSR 모델과 비교했을 때, coupled HMM이 factorial HMM이나 기존의 다른 모델보다 인식 성능이 좋음을 보였다[18].

본 논문에서는 DBN 프레임워크를 이용하여 동영상 플레이어나 PowerPointTM를 제어할 수 있는 손 제스처 인식 방법을 제안하였다. 한 손 또는 두 손 제스처만을 정의하고 인식한 기존 연구[7,12,19]와 달리, 정의된 제스처 모델은 한 손 제스처뿐만 아니라 두 손 제스처 모두를 하나의 프레임워크로 인식한다. 주어진 비디오 영상에서 두 가지 피부색 모델을 적용하여 영상 내에서 피부 영역을 검출하고, Argyros 등[20]이 제안한 방법의 변형을 통해 각 손의 움직임을 추적한다. 제안한 방법은 손과 손, 손과 얼굴 간의 겹침 현상 또는 비선형적 움직임에서도 강인한 추적을 할 수 있다. 손의 움직임과 손과 손, 손과 얼굴 간의 상대적 위치 정보를 관측값으로 하여 DBN을 이용한 새로운 제스처 인식 모델을 제안하였다. 10가지 제스처에 대한 교차 검증 실험에서 최대 99.59%의 높은 인식 성능을 보였다.

논문의 구성은 다음과 같다. 2장에서는 시스템의 전체적인 구성에 대해 설명하고, 3장에서는 얼굴과 손 영역을 검출하고, 추적하는 방법을 설명한다. 손 동작 정의 및 특징 추출은 4장, DBN을 이용한 동작 인식 모델의 정의, 추론 및 학습에 대한 것은 5장에서 각각 살펴본다. 6장에서는 실험 결과를 보이고, 7장에서 결론을 내린다.

2. 시스템 구성

본 논문에서 제안하는 시스템의 전제적인 구성은 그림 1에 나타난 바와 같이 여러 기능 모듈들의 파이프라인으로 이루어진다. 각 모듈은 다음의 역할을 수행한다.

- 배경 제거: 입력 비디오 시퀀스의 각 프레임에 대하여 배경 이미지와의 차이를 계산하여 전경 이미지를 추출한다.
- 피부색 검출: 두 가지의 피부 색상 모델을 이용하여 피부색을 추출한다. 두 가지 모델은 YIQ 색상 모델과 Haar-like 얼굴 검출기를 통해 검출된 얼굴 영역의

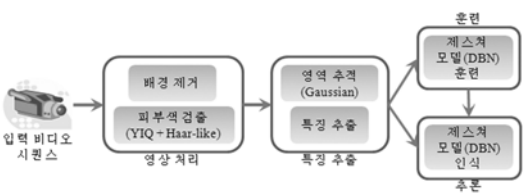


그림 1 시스템 구성

픽셀들을 이용하여 생성된 모델이다. 입력 영상의 픽셀들을 각 모델에 적용한 결과들을 결합하여 피부 영역을 검출한다.

- 영역 추적: 피부색 검출 단계에서 검출된 손과 얼굴의 각 영역들을 추적한다. 각 영역은 가우스(Gaussian) 함수로 모델링된다.
- 특징 추출: 손의 움직임을 이용하여 특징을 추출한다. 자세한 내용은 4장에서 설명한다.
- 제스처 모델 훈련: 각 손의 제스처 패턴에 대한 동적 특성들을 표현하는 DBN을 훈련한다.
- 제스처 모델 추론: 새로운 입력 비디오 시퀀스에 대한 인식을 수행한다.

3. 얼굴과 손 영역에 대한 검출 및 추적

얼굴과 손의 검출 및 추적은 손 동작 인식을 위해 필수적인 부분이며, 시스템의 성능에 크게 영향을 미친다. 검출 및 추적에 대한 본 논문에서의 해결 방안에 대해 살펴본다.

3.1 피부 영역 검출

피부색 검출은 빛의 조건에 크게 영향을 받는다. 많은 사람들이 피부색 검출에 대한 연구를 해 왔으며 지금도 활발히 연구하고 있지만, 다양한 환경 또는 조건의 변화에 강인한 방법은 아직까지 알려져 있지 않다[21].

본 논문에서는 피부 영역의 검출을 위하여 두 가지의 방법을 결합한다. 첫 번째는 여러 문헌에서 많이 사용되는 간단한 컬러 색상 모델로 YIQ 컬러 공간에서 피부색상의 범위를 미리 정해두고, 해당 범위 내에 속하는 픽셀들을 피부색으로 검출하는 방법이다. 두 번째는 Haar-like 얼굴 검출기[22]을 적용하여 검출된 얼굴 영역 내의 픽셀들을 이용하여 생성되는 피부색 모델로 현재의 빛 조건 및 사용자의 피부색 특성을 반영한다. 이는 Haar-like 얼굴 검출기가 빛 조건의 변화에서 정면을 향하고 있는 얼굴을 잘 찾는다는 것을 이용한 것이다. 영상 내에 있는 사용자의 얼굴을 검출하고, 검출된 얼굴 부분에서 너무 어둡거나 너무 밝지 않은 픽셀들을 이용하여 그 사용자에게 맞는 피부 색상 모델을 생성한다. 피부 색상 모델은 HSV 색상 모델에서의 색상값(hue)에 대한 히스토그램으로 정의한다[23]. $H = \{h(k), k=1,\cdots,K\}$를 K개의 레벨을 가지는 색상값에 대한 히스토그램이라 하고, 이산 함수 $h(k)=n_k$라 정의한다. 여기서, k는 k번째 레벨의 색상값을 의미하고, n_k는 얼굴 영역 내에서의 픽셀들 중 색상값이 k번째 레벨에 속하는 픽셀들의 수를 의미한다. 이를 이용하여 확률 분포를 정의하기 위해 이산 함수를 다음과 같이 정규화한다.

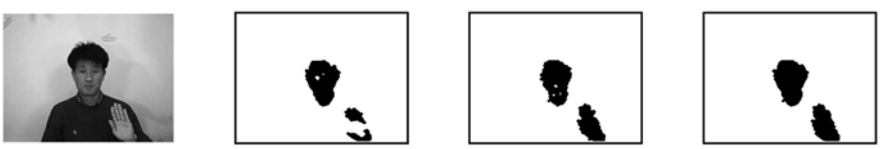
$$P_k = \frac{h(k)}{\sum_{l=1}^{K} h(l)}$$

위 식에서 k는 입력 픽셀이 속하는 색상값의 레벨을 의미한다. 그림 2는 서로 다른 모델을 적용한 피부색 검출 결과를 보이고 있다. 그림 2(b)는 YIQ 색상 모델을 이용한 것이고, 그림 2(c)는 얼굴색으로 만든 색상 모델만을 이용한 것이며, 그림 2(d)는 두 모델의 결과들에 대해 OR 연산을 적용한 결과이다.

3.2 영역 추적

비디오 영상에서의 손 추적은 개념적으로 간단하긴 하지만 얼굴 또는 다른 손과의 겹침 현상 때문에 쉽지 않은 문제이다. 본 논문에서는 개개의 손 및 얼굴에 해당하는 각 영역을 하나의 가우스 함수로 표현하여 추적하는 Argyros 등[20]의 방법을 이용한다. Argyros 등의 방법은 가우스 분포에서의 평균으로 얼굴과 손의 위치를 나타내며, 이전 두 프레임에서의 변화율(velocity)을 이용하여 현재 프레임에서의 손과 얼굴의 위치를 예측한다. 손과 얼굴의 위치 예측을 위해 이전 두 프레임에서의 변화율을 이용하는 방법은 손의 움직임이 일정한 속도를 가질 때 비교적 정확한 추적이 가능하지만 그렇지 못한 경우에는 오차를 범하게 되고, 추적에 실패하게 된다. 본 논문에서는 기존 방법과는 달리 손과 얼굴의 위치를 예측하기 위해 광류(optical flow)를 사용한다. 광류는 이전 프레임에서 현재 프레임으로의 손과 얼굴의 움직임에 대한 정보를 직접적으로 제공해주므로 추적하는 영역이 비선형적으로 움직이는 경우에도 정확한 추적을 한다.

손과 얼굴 영역을 추적하는 알고리즘[20]은 다음과 같다. 각 영역을 가우스 모델에 대한 타원 $g=(c_x, c_y, \alpha, \beta, \theta)$으로 표현한다고 하자. 여기서 (c_x, c_y)는 가우스 분포의 중심, (α, β)는 공분산 분석을 통해 계산된 주성



(a) (b) (c) (d)

그림 2 피부색 검출 방법 비교: (a) 입력 영상, (b) YIQ를 이용한 피부색 검출, (c) 얼굴 색상 분포를 이용한 피부색 검출, (d) (b)+(c)

분축들(제1주성분, 제2주성분), 그리고 θ는 입력 영상의 가로축과 제 1주성분축 사이의 각도를 의미한다. 각 영역과 이를 표현하는 가우스 모델 사이의 매핑은 다음과 같이 3가지 경우를 고려해 볼 수 있다.

- 새로운 영역이 화면상에 나타난 경우: 해당 영역의 픽셀들의 (x, y) 좌표값을 이용하여 새로운 가우스 모델을 생성
- 추적하던 영역이 화면에서 사라지는 경우: 이 영역을 표현하던 가우스 모델을 제거
- 각 영역의 추적 시 하나 이상의 영역들이 겹치는 경우: 두 가지 규칙을 적용하여 가우스 모델을 갱신

세 번째의 경우, 각 영역에 속한 픽셀들과 가우스 모델들과의 Mahalanobis 거리를 다음과 같이 계산하여 이를 가우스 모델의 갱신에 이용한다.

$$D(p,g)=\left(\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\cdot\begin{bmatrix}\frac{x-c_x}{\alpha}\\ \frac{y-c_y}{\beta}\end{bmatrix}\right)^2 \quad \text{where, } \forall p\in B_i, \forall g\in G$$

수식에서 B_i는 피부색 영역, p는 B_i에 포함된 픽셀들, G는 현재까지 생성된 가우스 모델의 집합을 의미한다. 영역내의 픽셀들과 가우스 모델과의 거리를 이용하여 각 영역과 가우스 모델들 사이에서의 매핑을 다음 규칙에 따라 결정한다.

$$\text{규칙 1: } R_1=\{p\in B \mid D(p,g)\le 1\}$$

$$\text{규칙 2: } R_2=\left\{p\in B \mid g=\arg\min_{g\in G}\{D(p,g)\mid D(p,g)>1\}\right\}$$

규칙 1은 픽셀이 가우스 모델에 대한 타원의 내부에 있는 경우로 픽셀이 가우스 모델의 갱신에 이용된다는 것을 의미하고, 규칙 2는 픽셀이 타원의 외부에 있는 경우로 가장 가까운 거리에 있는 가우스 모델의 갱신에 이용된다는 것을 의미한다. 가우스 모델과 각 영역과의 매핑 전에 가우스 모델의 평균 위치를 예측할 필요가 있는데, 이는 이전 프레임에서의 피부 영역과 현재 프레임에서의 피부 영역 사이에서의 광류값을 계산하여 이들의 평균값, $\mathbf{v}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{f}(i)$을 이용한다. 여기서 $\mathbf{f}(i)=\left[f_x(i), f_y(i)\right]^T$는 i번째 픽셀의 광류, N은 영역 내에서 검출된 광류의 개수를 의미한다. 현재 프레임에서의 가우스 모델의 중심을 식 (1)과 같이 이전 프레임에서의 위치에서 광류값들의 x방향과 y방향 각각에 대한 평균값만큼 이동한 위치로 예측한다.

$$g'=\left(c_x+\mathbf{v}_x, c_y+\mathbf{v}_y, \alpha, \beta, \theta\right)$$

식에서 g'는 현재 프레임에서의 영역 위치를 예측한 가우스 모델이다. 이는 이전 프레임에서 현재 프레임으로의 움직임에 대한 정보를 현재 프레임에서의 추적에 즉시 이용하므로 이전 두 프레임에서의 변화율을 이용한 방법보다 더욱 정확하고, 손의 움직임 속도 변화에도 민감하여 정확한 추적이 가능하게 한다.

그림 3은 이전 두 프레임에서의 변화율을 이용한 경우와 광류를 이용한 경우에 대한 결과를 비교하여 보여준다. 그림 3(b)는 이웃한 두 프레임에서의 손의 움직임

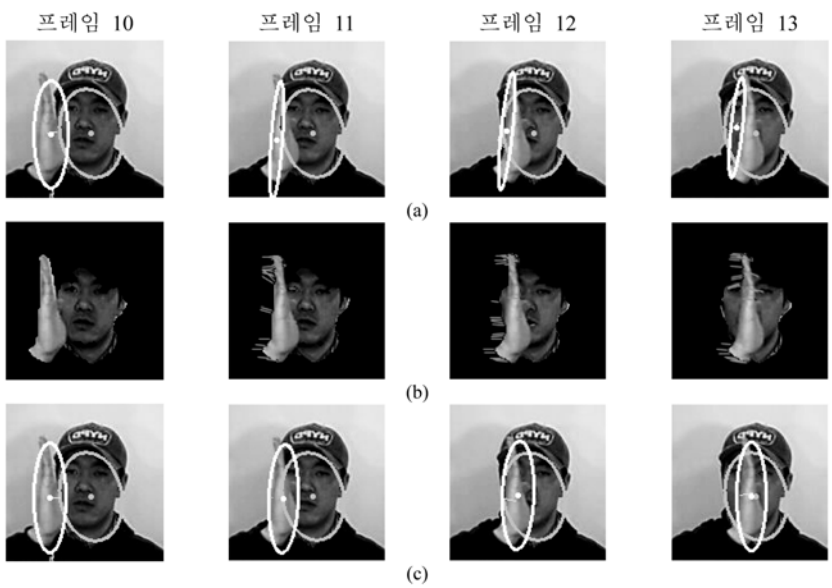


그림 3 변화율을 이용한 예측 방법과 광류를 이용한 예측 방법의 결과 비교: (a) 이전 두 프레임에서의 변화율을 이용한 예측, (b) 이전 프레임과 현재 프레임 사이에서의 광류, (c) (b)의 광류를 이용한 예측

에 대한 광류를 보여주고 있다. 프레임 10에서의 광류에 나타난 것과 같이 프레임 9와 프레임 10사이에 손의 움직임이 없었기 때문에 그림 3(a)의 변화율을 이용한 방법은 프레임 10과 프레임 11에서도 큰 움직임이 없을 것이라는 예측을 했으나, 실제로는 움직임이 나타나 오차가 생기게 된다. 이러한 오차의 누적으로 연속된 프레임에서 손의 정확한 위치를 찾지 못하여 프레임 13에서는 추적에 실패하게 된다. 그러나, 광류를 이용한 그림 3(c)는 프레임 10 이전에서의 손의 움직임과 상관없이 그림 3(b)에 나타난 바와 같이 직전 프레임과 현재 프레임간의 광류로부터 움직임에 대한 정보를 제공받기 때문에 추적에 성공하게 된다.

4. 동작 정의 및 특징 추출

4.1 동작 정의

본 논문에서는 미디어 플레이어 또는 PowerPoint™ 제어를 위한 10가지의 동작들을 그림 4와 같이 정의한다. 정의된 동작들을 5가지의 한 손 동작과 5가지의 두 손 동작을 포함한다. 그림에서 검은색 점은 손의 시작 위치, 화살표는 손의 이동 궤적을 의미한다.

4.2 특징 추출

손 동작의 인식에 있어서 가장 중요한 정보는 손의 움직임 정보이다. 움직임은 시간과 공간상에서 손의 이동 궤적이나 좌표 벡터의 시퀀스 $\mathbf{x}_t$, t=1, …,로 표현될 수 있다. 시간 t에서의 손의 위치 $\mathbf{x}_t$는 그림 3에 보여진 바와 같이 가우스 분포에서의 평균으로 표현된다. 연속한 프레임에서의 평균 변화로 손의 지역적 움직임 벡터를 표현한다. 움직임의 궤적은 지역적 움직임 벡터의 시퀀스로 표현될 수 있으며, 움직임 벡터는 그림 5(a)와

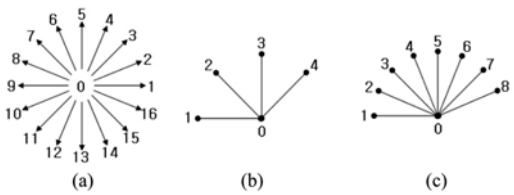


그림 5 특징 추출: (a) 17 방향 코드, (b) 5가지 상대적 위치, (c) 9가지 상대적 위치

같이 17방향 코드로 양자화하여 표현한다. 그림에서 가운데 위치한 '0'은 움직임이 없음을 의미한다. 매 프레임마다 두 손의 움직임에 대한 방향 코드가 생성된다.

손 동작 인식에 있어 각 손의 움직임을 독립적으로 고려할 경우, 손의 움직임에 대한 방향 정보만으로는 서로 다른 동작 간에 모호성이 생길 수 있다. 예를 들어, 영상 내에 두 개의 손이 모두 나타나 있고, 그림 6(a)의 '빨리 감기'나 그림 6(b)의 '뒤로 감기'와 같은 한 손 제스처를 취한다고 가정하자. 이들 제스처에 대한 방향 코드 시퀀스는 그림 6(c)의 '마지막 프레임으로 이동'이나 그림 6(d)의 '첫 프레임으로 이동' 제스처의 방향 코드 시퀀스와 동일한 형태를 가진다. 방향 코드를 이용한 이들 제스처를 단순화 형태가 그림 6(e)와 그림 6(f)에 각각 나타나 있다. 그림에서 채워진 원은 손의 시작 위치를 나타내고, 점선으로 된 원은 손의 마지막 위치를 나타내며, 화살표는 손의 이동 경로를 의미한다. 방향 코드 만을 이용했을 때 발생하는 이들 제스처 간의 모호성을 제거하기 위해 두 손의 상대적 위치 정보(그림 5(b))와 얼굴과 각 손의 상대적 위치 정보(그림 5(c))에 대한 두 가지의 특징을 추가로 이용한다. 두 경우 모두

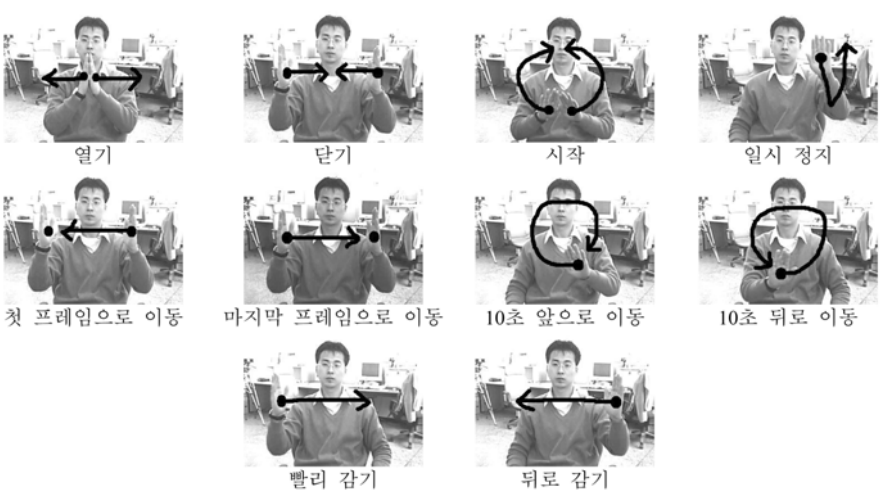
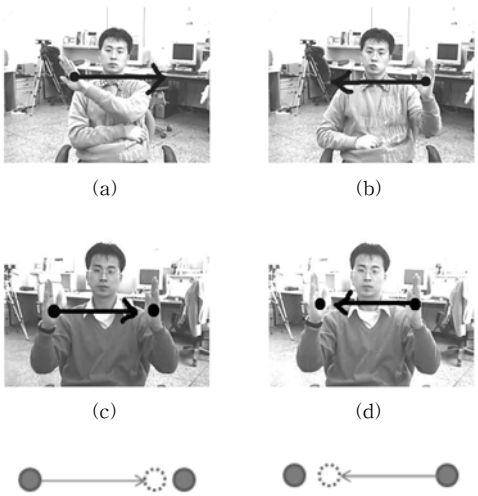


그림 4 미디어 플레이어 제어를 위한 손 동작 정의



(a) (b) (c) (d) (e) (f)

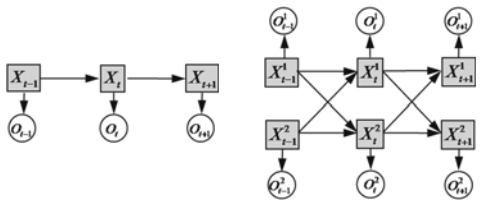
그림 6 방향 코드만을 이용한 경우에 발생하는 모호성: (a) '빨리 감기' 제스처, (b) '뒤로 감기'제스처, (c) '마지막 프레임으로 이동' 제스처, (d) '첫 프레임으로 이동' 제스처, (e) (a)와 (c)의 방향 코드를 이용한 제스처의 단순화된 형태, (f) (b)와 (d)의 방향 코드를 이용한 제스처의 단순화된 형태

'0'은 손과 손 또는 손과 얼굴이 겹쳐져 있음을 의미한다. 이들 정보가 무의식적인 작은 움직임 또는 영상 처리 과정에서의 오차에 의한 방향 코드의 변화로 인한 인식 오류를 줄여줄 수 있다는 것은 6장의 실험에서 보인다.

5. 인식 모델

5.1 동적 베이스망

은닉 마르코프 모델(HMM)[24]은 시계열 데이타를 모델링 하는데 있어서 유용한 도구로 널리 이용되고 있으나, 그 능력은 하나의 이산 상태 변수로 표현할 수 있는 상태 공간으로 제한되어 있다. 예를 들어, 두 개 이상의 은닉 노드에 대한 마르코프 과정을 그림 7(a)의 표준 HMM에서와 같이 하나의 은닉 노드로 표현한다고 가정하자. 그림 7에서 회색의 사각형은 은닉 노드(X)를 나타내고, 흰색의 원은 관측 노드(O)를 의미하며, 화살표는 이들 노드간의 의존성을 표시한다. 이때, 각 마르코프 과정에서의 노드들이 가질 수 있는 상태들을 모두 곱한 수만큼의 상태수를 표준 HMM에서는 하나의 상태 노드로 표현할 수 있어야 하므로 상태 공간의 크기가 지수적으로 증가하게 된다. 상태 공간의 크기가 급격히 증가함에 따라 신뢰할 수 있는 모델의 파라미터 값을 결정하기 위해 많은 양의 훈련 데이타를 필요로 한다.



(a) (b)

그림 7 그래프 모델로 표현한 표준 HMM와 coupled HMM: (a) 표준 HMM, (b) coupled HMM

Coupled HMM(CHMM)[8]은 두 개의 독립된 과정(process) 간의 상호 작용을 모델링하기 위해 표준 HMM에 변형을 가한 모델로 그림 7(b)에 나타난 바와 같이 두 개의 HMM에서 상태 변수들이 쌍을 이루고 있다. 이 모델은 간단한 상호작용 과정들을 모델링하는데 유용하긴 하나 각 HMM이 공유해야 하는 정보를 공유하지 못하고, 여러 개의 과정으로 확장 되는 경우 노드간에 이루는 쌍의 수가 증가하게 되어 계산량이 지수적으로 증가하게 된다. 또한 새로운 정보 또는 특징의 추가가 어렵다는 단점을 가진다.

DBN은 그래프 모델의 하나인 베이스망(BN)에 시간 정보가 추가된 모델로 시계열 데이타를 모델링하기에 유용한 도구이며, 최근 많은 관심을 받고 있다. BN은 사이클이 없는 방향 그래프로 각 노드는 하나의 확률 변수를 나타내고, 방향성을 가진 에지는 변수들간의 의존성을 표현한다. BN은 변수들간의 조건부 독립성 또는 d-분리[25]를 이용하여 변수들의 결합 확률 분포를 효율적으로 표현 및 계산한다. 조건부 독립성은 모델의 구조와 해당 모델에 대한 추론 및 학습을 수행하는데 필요한 계산을 간단하게 한다. BN과 같은 그래프 모델에서는 확률 변수들 간의 조건부 독립성을 수식적 분석 없이 그래프 상에 나타난 노드들의 관계로부터 직접적으로 확인할 수 있다는 장점을 가진다. DBN은 BN의 노드들에 대해 t-1 시간의 노드값이 t 시간에서의 노드값에 영향을 준다는 마르코프 과정을 적용하여 노드들 간의 시간적 의존 관계를 표현한 것으로, BN에서의 추론 알고리즘을 그대로 적용할 수 있다.

5.2 제안하는 양손 제스처 모델

4장에서 정의된 10가지 제스처는 한 손 제스처 뿐만 아니라 두 손 제스처도 포함하고 있다.두 손 제스처의 경우 오른손과 왼손의 움직임은 서로 연관성을 가지며 이런 동작들은 coupled HMM으로 모델링 될 수도 있지만, 본 논문에서 3개의 은닉 노드와 5개의 관측 노드를 가지는 새로운 모델을 제안한다. 은닉 노드 X^1, X^2는 오른손과 왼손의 움직임을 모델링하며, 은닉 노드들은 각각 방향 코드와 얼굴과의 상대적 위치 정보를 관

(a) (b)

그림 8 각 손을 모델링하는 노드들(X^1, X^2)과 새로운 노드(X^3)와의 관계 표현: (a) X^3에서 X^1, X^2로의 연결, (b) X^1, X^2에서 X^3으로의 연결

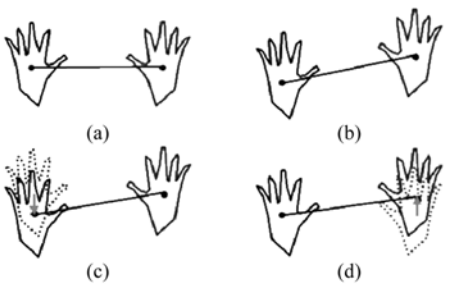


그림 9 두 손의 상대적 위치 변화: (a) 초기 상대적 위치, (b) 변화된 상대적 위치, (c) 왼쪽 손이 내려간 경우, (d) 오른쪽 손이 올라간 경우

측값으로 가진다. 은닉 노드 X^3는 유사한 동작들 간의 모호성을 제거하기 위한 것으로 두 손의 상대적 위치 변화를 모델링한다.

이들 은닉 노드들 간의 관계에 대해서는 그림 8과 같이 두 가지의 경우를 생각 해 볼 수 있다. 그림 8(a)는 두 손의 상대적 위치를 모델링하는 은닉 노드 X^3가 각각의 손을 모델링하는 노드 X^1과 X^2의 조건이 된다는 것을 나타내고 있으며, 그림 8(b)는 각각의 손을 모델링하는 노드 X^1과 X^2가 손의 상대적 위치를 모델링하는 은닉 노드 X^3의 조건이 됨을 표현하고 있다.

손의 상대적 위치가 그림 9(a)에서 그림 9(b)로 바뀌는 경우를 살펴보자. 이때, 만약 왼손이 아래로 움직였다면 오른손은 위로 움직이거나 움직임이 없어야 한다. 반대로, 오른손이 위로 움직인 경우, 왼손은 아래로 움직이거나 움직임이 없어야 한다. 이는 두 손의 상대적 위치에 대한 은닉 노드 X^3의 값이 주어진 경우, 은닉 노드 X^1, X^2는 조건부 의존(conditional dependent)의 관계가 성립함을 의미한다. 한편, 두 손의 상대적 위치에 대한 정보가 주어지지 않은 경우, 각각의 손은 다른 손의 움직임과 상관없이 독립적인 움직임을 가질 수 있게 되며, 두 손의 상대적 위치는 각 손의 움직임의 결과에 따라 결정된다. 즉, 은닉 노드 X^3의 값이 주어지지 않은 경우에 은닉 노드 X^1과 X^2는 주변 독립(marginal independent)의 관계를 가진다. 그림 8(b)는 이러한 관계를 모델링하고 있다. 그러나 그림 8(a)는 이와는 반대의 개념을 모델링하고 있다. 두 손의 상대적 위치에 대한 은닉 노드 X^3의 값이 주어진 경우에 은닉 노드 X^1, X^2는 조건부 독립(conditional independent)의 관계를 가지고, 그렇지 않은 경우에는 주변 의존(marginal dependent)의 관계를 가진다. 따라서, 그림 8(b)를 제안하는 손 제스처 모델에 반영한다.

시간의 흐름에 따른 은닉 노드들 간의 관계에 대해서 1차 마르코프 과정을 가정하여 그림 10과 같은 손 제스처 모델을 제안한다. 그림 10에서 회색 노드들은 은닉 노드들을 나타내고, 흰색의 노드들은 관측 노드들을 나타낸다. 두 손의 상대적 위치를 모델링하는 노드 X^3의 시간적 의존 관계를 나타내는 $\cdots \rightarrow X_{t-1}^3 \rightarrow X_t^3 \rightarrow X_{t+1}^3 \rightarrow \cdots$은 CHMM에서 완전히 연결되어 있던 두 노드 X^1과 X^2의 높은 의존성을 완화시켜 서로간의 의존성을 간접적으로 표현하며, X^1과 X^2가 가질 수 있는 공유 정보를 포함한다. 따라서, 시간 t-1에서의 각 손의 움직임 정보와 시간 t에서의 각 손에 대한 움직임의 정보가 노드 X_{t-1}^3과 X_t^3에 의해 간접적으로 연결되므로 CHMM에서 완전히 연결되어 있던 이들 두 노드 사이의 연결은 제거될 수 있다.

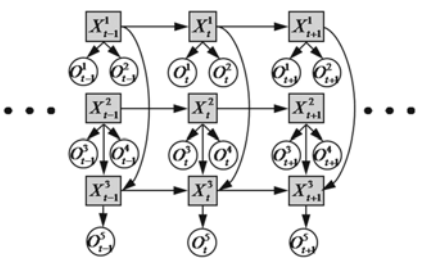


그림 10 손 동작 인식을 위한 DBN

5.3 추 론

BN은 변수들간의 조건부 독립 관계를 그래프 상에서 쉽게 알 수 있도록 하며, 결합 확률 분포를 효율적이고 쉽게 구할 수 있도록 한다는 장점을 가진다. BN에서의 결합 확률 분포는 d-분리에 따른 조건부 확률 분포의 곱으로 간단히 표현될 수 있다는 것을 DBN에 그대로 적용할 수 있다. 즉, 시간 t-1의 BN이 주어졌을 때, 시간 t의 BN은 시간 t-1이전의 모든 BN과 조건부 독립이 되므로 시간 t에서의 BN에 대한 확률 분포 계산을 위해서는 시간 t-1의 BN만 고려하면 된다. 따라서, 제안하는 DBN에서의 결합 확률 분포는 식 (2)와 같이 계산될 수 있다.

$$P(X_{1:T}^{1:3}, O_{1:T}^{1:5}) = P(X_1^1)P(X_1^2)P(X_1^3 \mid X_1^1, X_1^2)$$

$$\times\prod_{t=2}^{T} P(X_t^1 \mid X_{t-1}^1)P(X_t^2 \mid X_{t-1}^2)P(X_t^3 \mid X_{t-1}^3, X_t^1, X_t^2)$$
$$\times\prod_{t=1}^{T} P(O_t^1, O_t^2 \mid X_t^1)P(O_t^3, O_t^4 \mid X_t^2)P(O_t^5 \mid X_t^3)$$
$$\text{where, } X_{1:T}^{1:3} = \begin{bmatrix} X_1^1 \\ X_1^2 \\ X_1^3 \end{bmatrix} \cdots \begin{bmatrix} X_T^1 \\ X_T^2 \\ X_T^3 \end{bmatrix} \text{ and } O_{1:T}^{1:5} = \begin{bmatrix} O_1^1 \\ \vdots \\ O_1^5 \end{bmatrix} \cdots \begin{bmatrix} O_T^1 \\ \vdots \\ O_T^5 \end{bmatrix} \quad (2)$$

DBN에서의 추론을 위해 interface 알고리즘[26]을 이용한다. V_t를 시간 t에서의 노드들의 집합, E를 DBN에서의 노드 간의 에지(edge) 집합이라 할 때, $E^{tmp}(t)$를 식 (3)과 같이 이웃한 두 시간에서의 BN을 연결하는 노드 간의 에지 집합이라 정의하자.

$$E^{tmp}(t) = \{(u,v) \in E \mid u \in V_{t-1}, v \in V_t\} \quad (3)$$

Interface 알고리즘은 DBN에서 이웃하는 두 시간에서의 BN(2 Time-slice BN: 2TBN)을 고려했을 때, 시간 t-1에서 시간 t로 연결된 화살표를 가진 노드들의 집합은 과거(시간 t-1이전의 모든 BN)와 미래(시간 t이후의 모든 BN)를 d-분리 하기에 충분하다는 것을 이용한 알고리즘이다. 이때, 과거와 미래를 d-분리 해주는 노드들의 집합을 '*interface*'라 한다. $ch(v)$를 노드 v의 자식 노드들의 집합이라 했을 때, interface는 다음과 같이 정의된다.

$$I_t^{\rightarrow} \triangleq \{u \in V_t \mid (u,v) \in E^{tmp}(t+1),\ v \in V_{t+1}\}$$
$$I_t^{\leftarrow} \triangleq v \in V_t \mid (u,v) \in E^{tmp}(t) \text{ or } \exists w \in ch(v) : (u,w) \in E^{tmp}(t), u \in V_{t-1}$$

전진 interface는 현재 시간 t의 BN에서 다음 시간 t+1의 BN으로 연결된 화살표를 가진 노드들의 집합이고, 후진 interface는 시간 t-1에서 현재 시간 t로의 화살표를 가진 노드들과 그 노드들의 모든 자식 노드들의 집합을 의미한다. 따라서 2TBN에서 interface 노드들의 값이 주어지면 이는 과거와 미래를 d-분리하기에 충분하며, 2TBN에서 interface에 속하지 않는 노드들을 제거함으로써 1.5TBN으로 줄일 수 있다. 1.5TBN을 접합(junction) 트리로 변형하여 추론 알고리즘을 수행한다. Interface 알고리즘은 DBN을 접합(junction) 트리로 변환 할 때, interface에 속하는 노드들을 포함하는 clique 노드가 반드시 존재해야 한다는 조건을 가진다. 그래프에서 clique는 모든 노드들의 쌍이 에지를 가지고 있는 노드들의 집합을 말한다. 접합 트리를 구성한 뒤에는 접합 트리 알고리즘(Junction-Tree Algorithm: JTA)[27]을 적용할 수 있다.

JTA는 그래프 모델에서 정확한(exact) 값을 추론하는 알고리즘이다. 접합 트리 알고리즘을 적용하기 위해서는 방향성을 가진 그래프를 방향성이 없는 그래프로의 변환을 거친 뒤 접합 트리를 구성해야 한다. 제안하는 DBN 모델에 대한 접합 트리 생성 과정이 그림 11에 나타나 있다. 그림에서 clique 내에 회색의 음영으로 표현된 것은 interface 노드들을 표현한 것이다. 방향성이 없는 그래프로의 변환은 노드간의 조건부 독립 특성이 사라지도록 하는데 이로 인해 그림 11(c)의 교화(moralization)와 그림 11(d)의 삼각화(triangulation) 과정이 필요하다[27]. 이웃한 두 clique C_1, C_2 사이에는 두 clique에 공통적으로 포함된 노드들이 존재하는데 이들 노드들의 집합을 분리자(separator)를 S_{12}라 하자. 예를 들어, 그림 11에서 두 clique $X_{t-1}^1, X_{t-1}^2, X_{t-1}^3, X_t^1$과 $X_{t-1}^2, X_{t-1}^3, X_t^1, X_t^2$에 대한 분리자는 $X_{t-1}^2, X_{t-1}^3, X_t^1$이 된다. 그림 11에서 분리자는 점선으로 된 사각형으로 표시되어 있다. 접합 트리 알고리즘은 메시지-전달(message-passing) 프로토콜을 이용한다. 즉, 어떤 clique가

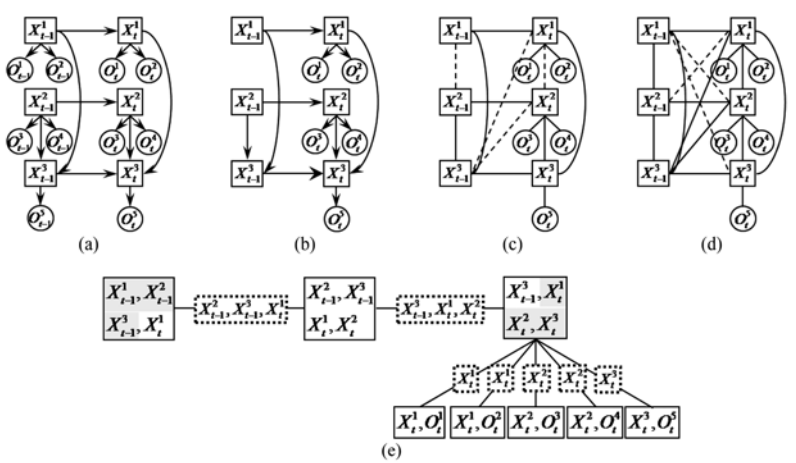


그림 11 Interface 알고리즘을 위한 접합 트리의 구성 과정: (a) 2TBN, (b) 1.5TBN, (c) 교화된 그래프, (d) 삼각화된 그래프, (e) 제안된 DBN에서의 2TBN에 대한 접합 트리

다른 clique로 메시지를 전달하기 위해서는 전달하고자 하는 노드를 제외한 나머지 모든 노드로부터 메시지를 전달 받은 후에 메시지를 전달할 수 있다. $f(C_1)$과 $f(S_{12})$를 각각 clique C_1과 분리자 S_{12}에 대한 potential 함수라 했을 때, Clique C_1에서 C_2로 메시지를 전달하는 과정은 다음과 같다. 우선 clique C_1은 분리자 S_{12}로 메시지를 전달한다. C_1의 potential 함수에서 S_{12}에 포함된 노드들을 제외한 모든 노드들($C_1 \backslash S_{12}$)이 가질 수 있는 모든 값들을 더해 S_{12}의 새로운 potential 함수 $f'(S_{12})$로 지정한다.

$$f'(S_{12}) = \sum_{C_1 \backslash S_{12}} f(C_1)$$

그리고 분리자 S_{12}는 C_1으로부터 메시지를 받은 후에 C_2로 다음의 연산을 통해 메시지를 전달하여 C_2의 potential 함수를 갱신한다.

$$f'(C_2) = f(C_2)\frac{f'(S_{12})}{f(S_{12})}$$

DBN에서는 이와 같은 방식으로 반복적으로 메시지를 전달하여 추론을 수행한다.

5.4 학습

DBN에서의 학습은 훈련 데이타($\mathbf{O}$)가 주어졌을 때, 식 (4)와 같이 주어진 관측 데이타에 대한 DBN의 우도를 최대로 하는 파라미터의 값을 결정한다.

$$\hat{\theta} = \arg\max_{\theta} P(\mathbf{O} \mid \theta) \tag{4}$$

식에서 파라미터 θ는 HMM을 포함한 모든 상태 공간 확률 모델에서와 마찬가지로, 은닉 노드의 초기 확률 분포(π), 은닉 노드의 상태간의 상태 전이 확률 분포(A), 은닉 노드의 임의의 상태에서의 관측값 출력 분포(B)포함한다. 즉, $\theta = (\pi, A, B)$이다. 하지만, 은닉 상태를 하나의 확률 변수로 표현하는 HMM과는 달리, DBN은 여러 개의 확률 변수들로 은닉 상태를 표현한다.

본 논문에서 제안한 DBN은 은닉 노드들을 포함하고 있으므로 파라미터 훈련은 EM 알고리즘[28]을 적용한다. 관측 데이타 $\mathbf{O}$가 주어졌을 때, 모델의 로그-우도값은 다음과 같이 계산된다.

$$\begin{aligned} l(\theta, \mathbf{X}) &= \log \sum_{\mathbf{X}} P(\mathbf{O}, \mathbf{X} \mid \theta) = \log \sum_{\mathbf{X}} Q(\mathbf{X}) \frac{P(\mathbf{O}, \mathbf{X} \mid \theta)}{Q(\mathbf{X})} \\ &\geq \sum_{\mathbf{X}} Q(\mathbf{X}) \log \frac{P(O, \mathbf{X} \mid \theta)}{Q(\mathbf{X})} \\ &= \sum_{\mathbf{X}} Q(\mathbf{X}) \log P(O, \mathbf{X} \mid \theta) - \sum_{\mathbf{X}} Q(\mathbf{X}) \log Q(\mathbf{X}) = F(Q, \theta) \end{aligned} \tag{5}$$

X는 은닉 노드들의 집합, Q(X)는 은닉 노드들의 확률 분포를 의미한다. 간단한 수식적 조작을 통해 다음과 같은 식을 얻을 수 있다.

$$\begin{aligned} l(\theta, \mathbf{X}) - F(Q, \theta) &= \sum_{\mathbf{X}} Q(\mathbf{X}) \log Q(\mathbf{X}) - \sum_{\mathbf{X}} Q(\mathbf{X}) \log P(O, \mathbf{X} \mid \theta) \\ &= KL[Q \parallel P(X \mid O, \theta)] \geq 0 \end{aligned} \tag{6}$$

여기서 KL은 Kullback-Leibler divergence를 의미한다. 식 (6)으로부터 $Q(\mathbf{X}) = P(\mathbf{X}|O, \theta)$일 때, 최적값을 가진다는 것을 알 수 있다. 식 (5)의 마지막 줄의 두 번째 항목 $-\Sigma_X Q(\mathbf{X}) \log Q(\mathbf{X})$는 $\mathbf{X}$에 대한 엔트로피로 파라미터 θ를 포함하지 않으므로 θ와는 독립적이다. 즉, 로그-우도값을 최대로 하는 것과는 관련이 없으므로 고려할 필요가 없다. 그리고, 첫 번째 항목의 $\Sigma_{\mathbf{X}} Q(\mathbf{X}) \log P(\mathbf{O},\mathbf{X}|\theta)$는 5.3절의 추론에서 설명한 바와 같이 DBN에서의 조건부 독립성, interface 알고리즘, 그리고 접합 트리 알고리즘을 이용하여 계산될 수 있다.

파라미터 θ의 값을 결정하기 위해서는 훈련 데이타로부터 각 노드에 대한 충분 통계량만 계산하면 되는데, 이들 충분 통계량은 기대값을 이용하여 수식화 될 수 있다. 기대값은 앞서 설명한대로 1.5TBN을 접합 트리로 변환한 후 접합 트리 알고리즘을 수행하여 개개의 노드에 대한 marginalization을 계산하여 얻을 수 있다. 제안하는 DBN에서의 파라미터 집합 (π, A, B)에 대한 갱신 방법은 다음과 같다.

$$\begin{aligned} \hat{\pi}_i^q &= E\left[X_1^q = i\right] = \frac{P(O^{1:5}, X_1^q = i \mid \theta)}{P(O^{1:5} \mid \theta)} \text{ where } q \in \{1, 2\} \\ &= \left(\text{시간 1에서 } X^q \text{가 상태 } i \text{로부터 전이가 일어난 수의 기대 비율}\right) \end{aligned} \tag{7}$$

$$\begin{aligned} \hat{\pi}_{ijk}^3 &= E\left[X_1^3 = k \mid X_1^1 = i, X_1^2 = j\right] = \frac{P(O^{1:5}, X_1^1 = i, X_1^2 = j, X_1^3 = k \mid \theta)}{P(O^{1:5}, X_1^1 = i, X_1^2 = j \mid \theta)} \\ &= \begin{pmatrix} \text{시간 1에서 } X^1 \text{과 } X^2 \text{가 각각 상태 } i \text{와 상태 } j \text{에 머무를 때} \\ X^3 \text{가 상태 } k \text{로부터 상태 전이가 일어난 수의 기대 비율} \end{pmatrix} \end{aligned} \tag{8}$$

$$\begin{aligned} \hat{A}_{ij}^1 &= \frac{E\left[X_t^1 = j \mid X_{t-1}^1 = i\right]}{E\left[X_{t-1}^1 = i\right]} = \frac{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_{t-1}^1 = i, X_t^1 = j \mid \theta)}{P(O^{1:5} \mid \theta)}}{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_{t-1}^1 = i \mid \theta)}{P(O^{1:5} \mid \theta)}} \\ &= \left(\text{상태 } i \text{에서 상태 } j \text{로 전이가 일어난 수의 기대 비율}\right) \end{aligned} \tag{9}$$

$$\hat{A}^2_{gh} = \frac{E\left[X_t^2 = h \mid X_{t-1}^2 = g\right]}{E\left[X_{t-1}^2 = g\right]} = \frac{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_{t-1}^2 = g, X_t^2 = h \mid \theta)}{P(O^{1:5} \mid \theta)}}{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_{t-1}^2 = g \mid \theta)}{P(O^{1:5} \mid \theta)}} \tag{10}$$

$= (\text{상태 } g\text{에서 상태 } h\text{로 전이가 일어난 수의 기대 비율})$

$$\hat{A}^3_{klmn} = \frac{E\left[X_t^3 = n \mid X_{t-1}^3 = m, X_t^1 = k, X_t^2 = l\right]}{E\left[X_{t-1}^3 = m, X_t^1 = k, X_t^2 = l\right]} = \frac{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_t^3 = n, X_{t-1}^3 = m, X_t^1 = k, X_t^2 = l \mid \theta)}{P(O^{1:5} \mid \theta)}}{\sum_{t=2}^{T} \frac{P(O^{1:5}, X_{t-1}^3 = m, X_t^1 = k, X_t^2 = l \mid \theta)}{P(O^{1:5} \mid \theta)}} \tag{11}$$

$= (X^3\text{이 } m\text{의 상태에 있으면서 상태 } triple\ \langle k,l,n\rangle\text{으로 전이한 기대 비율})$

$$\hat{B}_{iy} = \frac{E\left[O_t = y \mid X_t = i\right]}{E\left[X_t = i\right]} = \frac{\sum_{t=1}^{T} \frac{P(O = y, X_t = i \mid \theta)}{P(O^{1:5} \mid \theta)}}{\sum_{t=1}^{T} \frac{P(X_t = i \mid \theta)}{P(O^{1:5} \mid \theta)}} \text{ where } (O_t, X_t) \in \left\{ \begin{matrix} \left(O_t^1, X_t^1\right), \left(O_t^2, X_t^1\right), \left(O_t^3, X_t^2\right), \\ \left(O_t^4, X_t^2\right), \left(O_t^5, X_t^3\right) \end{matrix} \right\} \tag{12}$$

$= (\text{상태 i에서 관측 심볼 } y\text{가 관측된 기대 비율})$

위의 공식에서 $\hat{\pi}, \hat{A}, \hat{B}$ 은 각각 갱신된 파라미터 값을 의미한다. 파라미터 갱신에 필요한 기대값들은 접합 트리 알고리즘에서 해당 노드를 포함하고 있는 clique에서 해당 노드를 제외한 노드들이 가질 수 있는 값들을 모두 더해 제거(marginalization)함으로서 얻을 수 있다. EM 알고리즘은 다음과 같이 관측될 수 없는 노드들의 값을 추정하는 Expectation과 추정된 은닉 노드들의 값을 이용하여 식 (7)부터 식 (12)까지의 공식과 같이 파라미터의 값을 갱신하는 Maximization으로 구성된다.

$$\textbf{Expectation}: q^{t+1} \leftarrow \arg\max_{q} F\left(q^t, \theta^t\right)$$

$$\textbf{Maximization}: \theta^{t+1} \leftarrow \arg\max_{\theta} F\left(q^{t+1}, \theta^t\right)$$

학습 알고리즘은 이 두 단계를 수렴할 때까지 반복적으로 수행된다. 여기서, q^t는 t번째 반복에서 추정된 은닉 변수들의 값을 의미하고, θ^t는 t번째 반복에서의 DBN의 파라미터를 의미한다.

만약 훈련 데이타가 충분하지 못해 특정값에 대한 관측이 훈련 데이타에 포함되어 있지 않다면 해당값의 확률 분포는 0이 되고, 새로운 제스처의 입력에 대한 관측값에서 그 값이 관측될 경우, 모델 전체에 대한 확률은 0이 되게 된다. 이런 결과를 막기 위해 확률 분포에 대한 바닥치(floor smoothing)를 설정하고, 확률 분포를 재정규화 한다. 예를 들어, 은닉 노드 X^1에 대한 초기 상태 확률 분포의 경우 바닥치의 값을 v라 했을 때 다음과 같이 재정규화 할 수 있다.

$$P\left(X^1 = i\right) = \frac{E\left[X^1 = j\right] + v}{\sum_{j=1}^{L}\left(E\left[X^1 = j\right] + v\right)}$$

6. 실험 및 결과

6.1 실험 환경

실험 데이타는 소형 CMOS 카메라를 이용하여 획득하였고, 10 가지의 동작에 대해 7명이 서로 다른 날 각각 7번씩 촬영하여 총 490개의 비디오 데이타를 수집하였다. 초당 30프레임으로 촬영하였으며, 각 프레임의 크기는 320×240이고, 24비트 컬러 색상이다. 실험에 사용된 시스템은 Visual C++ 6.0과 Matlab 7.0으로 구현하였으며, Intel OpenCV 라이브러리 [29]와 Bayes Net Toolbox(BNT) [30]를 이용하였다.

6.2 영상 처리 및 특징 추출

시스템의 전체적 성능은 영상 처리를 통해 획득한 특징 벡터들의 양질 여부에 큰 영향을 받는다. 그림 12는 영상 처리의 과정 및 각 단계에서의 결과를 보여주고 있다. 그림 12(a)는 입력 프레임이며, 그림 12(b)-12(f)는 입력 영상에서 피부색 영역을 검출하는 단계를 보여주고 있다. 그림 12(b)는 배경 이미지와의 차이를 이진화 하여 획득한 전경 이미지이고, 그림 12(c)는 YIQ 색상 모델을 적용한 피부색 검출 결과를 보여주고 있다. 그림 12(d)는 Haar-like 얼굴 검출기를 이용하여 검출된 얼굴의 위치를 나타내고 있으며, 검출된 얼굴 영역에 있는 픽셀들의 RGB 컬러 색상을 HSV 컬러 색상으로 변환한 뒤, 색상값에 대한 히스토그램의 분포를 생성하고 이를 이용하여 검출된 피부 영역이 그림 12(e)에 나타나 있다. 그림 12(f)는 (b), (c), (e)를 결합하여 검출된 피부 영역들을 나타내고 있다. 그림 12(g)-12(j)는 검출된 피부 영역들의 추적 과정의 결과를 보인다. 그림 12(g)는 현재 프레임에서 검출된 피부 영역과 이전 프레임에서 검출된 피부 영역 사이에서 계산된 광류를 보여주고 있다. 계산된 광류에 대하여 x방향과 y방향에 대한 평균값을 이용하여 각 피부 영역의 분포를 표현하는

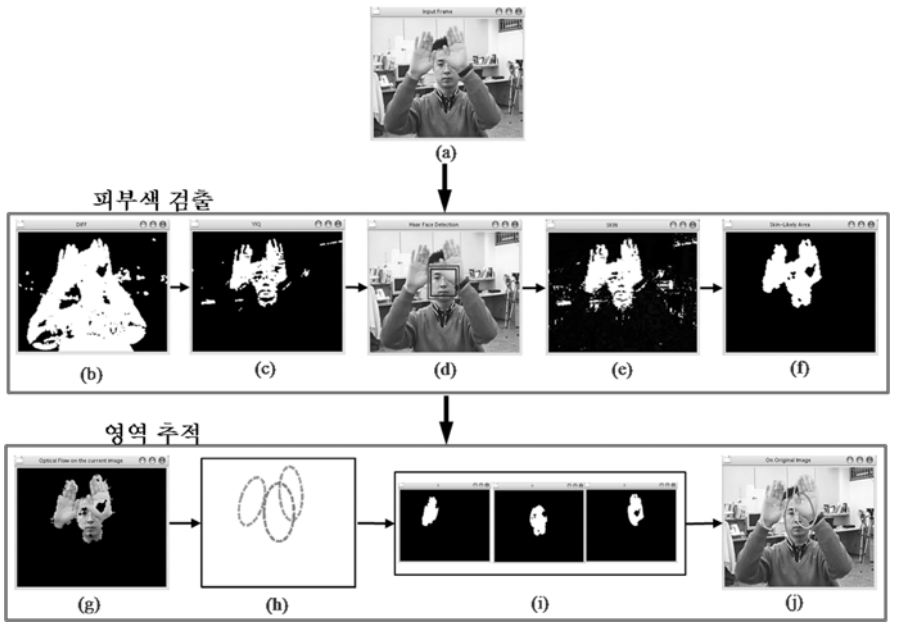


그림 12 영상 처리 과정 및 단계별 결과: (a) 입력 프레임, (b) 전경 픽셀 추출, (c) YIQ 색상 모델을 이용한 피부색 검출, (d) 얼굴 검출, (e) 얼굴 영역의 픽셀로 생성된 피부색 모델을 이용한 피부색 검출, (f) 검출된 피부 영역, (g) 연속된 프레임간의 광류, (h) 가우스 분포의 평균 위치 예측, (i) 영역 분할, (j) 가우스 분포 갱신

가우스 모델의 평균을 그림 12(h)와 같이 이동하여 현재 프레임에서의 손 및 얼굴에 대한 위치를 예측한다. 예측된 가우스 모델을 이용하여 Argyros 등[20]이 제안한 방법을 따라 그림 12(i)와 같이 손과 얼굴에 해당하는 영역들을 분리하고, 분리된 영역들의 픽셀들을 이용하여 가우스 모델의 평균 및 공분산에 대한 값들을 갱신하여 그림 12(j)의 결과를 얻는다.

6.3 인식(Isolated Recognition)

훈련 및 테스트 데이타의 부족으로 제스처 모델 성능은 7-fold 교차 검증 방법을 이용하여 평가하였다. 490개의 데이타에 대해 7개의 그룹을 형성하여 6개의 그룹으로 훈련하고 나머지 하나의 그룹으로 테스트를 하는 방식을 반복적으로 수행하였다. 훈련 및 테스트 데이타는 비디오 시퀀스에서 제스처에 해당하는 부분을 직접 선택하여 생성하였다. 총 7번의 교차 검증 실험에 대한 평균값을 인식률로 결정하였으며, 각각의 은닉 노드가 가질 수 있는 상태수는 동작의 복잡성을 고려하여 결정하였다. 입력 비디오 영상에 대한 손 제스처 동작 인식은 식 (13)과 같이 비디오 영상에서 추출한 특징값 ($O_{1:T}^{1:5}$)을 DBN의 관측 데이타로 하였을 때, 이들 관측 데이타에 대한 가장 큰 우도를 발생시키는 모델을 선택하였다. 수식에서 λ는 제스처 모델의 레이블을 의미하고, θ_λ는 DBN의 파라미터로 초기확률(π), 상태 전이확률(A), 상태별 출력 확률분포(B)를 포함한다.

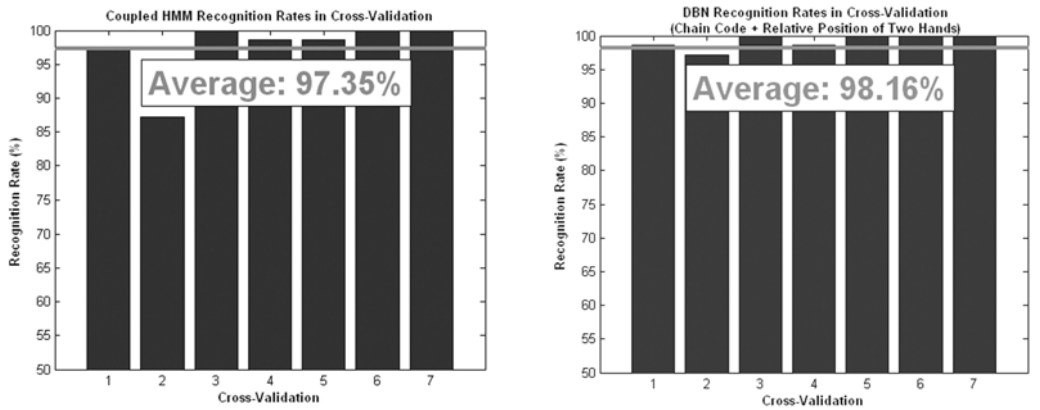
$$\hat{\lambda} = \arg\max_{\lambda} \left\{ P(O_{1:T}^{1:5} \mid \theta_\lambda) \right\} \tag{13}$$

6.3.1 Coupled HMM을 이용한 모델링

제안한 모델의 성능 평가를 위해 우선, 각각의 제스처를 coupled HMM으로 모델링하여 훈련 및 테스트를 해보았다. 한 손 제스처에 대해서 제스처를 취하지 않는 다른 손에 대한 움직임을 고려하지 않기 위해 균일(uniform) 분포를 지정하였다. 한 손 제스처에 대해서 이미지 상에 한 손만 나타나는 비디오 영상을 포함한 테스트에 평균 97.35%의 인식률을 얻었으며, 교차 검증 결과가 그림 13(a)에 나타나 있다. 그림에서 가로축은 전체 데이타에서 n번째 그룹 데이타를 테스트 데이타로 사용했음을 의미한다. 두 번째 그룹 데이타에서 상대적으로 낮은 인식률을 보였는데 이는 일부 프레임에서 갑작스러운 조명 변화로 인해 영상 처리 과정에서의 영역 검출 에러 및 추적 에러로 인한 것이었다. 특히 피부색 영역을 검출하는 과정에서 손 영역의 상당 부분의 픽셀들을 제대로 검출하지 못함으로 해서 각 영역을 모델링하는 가우스 분포의 평균값이 안정적이지 못한 결과를 보였으며, 그 결과 방향 코드에서의 급격한 변동을 초래하였다.

6.3.2 동적 베이스망을 이용한 모델링

손의 움직임을 표현하는 방향 코드뿐만 아니라, 두 손



(a) (b)

그림 13 Coupled HMM과 동적 베이스망으로 모델링한 인식 결과 비교: (a) Coupled HMM으로 모델링, (b) 동적 베이스망으로 모델링

의 상대적 위치 정보를 추가한 DBN을 이용한 인식에서는 그림 13(b)에 보여진 바와 같이 평균 98.16%의 성능을 보였으며, 이는 coupled HMM의 성능보다 조금 높은 결과이다.

한 손 제스처에 대해 두 손이 모두 이미지 상에 나타난 별도의 테스트 데이타에 대해 coupled HMM과 DBN의 성능을 비교해 보았다. 표 1에 나타난 바와 같이 6개의 테스트 데이타에 대해서 coupled HMM은 하나의 제스처만 정확하게 인식한 반면, DBN을 이용한 인식에서는 6개 모두를 정확하게 인식하였다. 이는 '빨리 감기', '마지막 프레임으로 이동', '뒤로 감기', '첫 프레임으로 이동' 제스처에 대해 새로이 추가된 두 손의 상대적 위치가 방향 코드만을 사용했을 때 발생하는 동작의 모호성을 해결하는 방안이 된다는 것을 보여준다.

DBN을 이용한 인식에서의 오인식은 '닫기', '마지막 프레임으로 이동', '첫 프레임으로 이동' 제스처에 대한 것들이었다. 오인식의 원인은 사용자가 고정되어 있어야 하는 손을 무의식적으로 조금씩 움직이거나 영상 처리

표 1 한 손 명령에 두 손이 영상에 나타난 경우의 인식: 정인식(○), 오인식(X), 오인식에 대한 인식 결과는 괄호 내에 표시

입력 비디오 영상	인식 결과	
	Coupled HMM	DBN
빨리 감기	X (마지막 프레임으로 이동)	○
일시 정지	○	○
10초 뒤로 이동	X (일시 정지)	○
10초 뒤로 이동	X (마지막 프레임으로 이동)	○
뒤로 감기	X (첫 프레임으로 이동)	○
10초 앞으로 이동	X (마지막 프레임으로 이동)	○
맞힌 수	1	6

표 2 DBN을 이용한 제스처 인식 결과

제스처	테스트 수	정인식 수	오인식 수	인식률 (%)
열기	49	49	0	100
닫기	49	49	0	100
시작	49	49	0	100
일시 정지	49	49	0	100
첫 프레임으로 이동	49	48	1	97.59
마지막 프레임으로 이동	49	48	1	97.59
10초 앞으로 이동	49	49	0	100
10초 뒤로 이동	49	49	0	100
빨리 감기	49	49	0	100
뒤로 감기	49	49	0	100
맞힌 수	490	488	2	99.59

의 피부색 검출에서의 에러로 인한 것들이었다. 이들 문제는 얼굴과 손의 상대적 위치 정보를 DBN에 추가하여 해결할 수 있었으며, 그 결과 평균 99.59%의 인식률을 얻었다. 이에 대한 결과가 표 2에 나타나 있다. 표에서 각각의 수는 테스트한 비디오 시퀀스의 수와 그들의 인식률을 의미한다.

6.3.3 은닉 상태 디코딩

인식 성능이 높다는 사실 만으로는 모델의 내부적 동작에 대한 것을 설명해 줄 수 없다. 그러나 은닉 노드에 대한 정보를 확률적으로 추정해 볼 수는 있다. HMM에서의 Viterbi 알고리즘이 이와 같은 역할을 수행한다. 관측 데이타가 주어졌을 때, 이들 관측 데이타에 대해 제안된 모델이 각 제스처를 잘 모델링하고 있는지를 확인하기 위해 은닉 노드들의 최적 상태 시퀀스를 디코딩하였다. 여기서, 최적 상태 시퀀스는 관측 데이타 $O_{1:T}^{1:5}$에 대해 데이타를 모델링하고 있는 DBN에서 시간 1부터

시간 T까지의 상태 노드들의 모든 경로들 중에서 최대 우도를 출력하는 하나의 상태 노드들의 경로를 의미한다. 최대 우도를 출력하는 상태 노드들의 시퀀스를 결정하기 위해서는 접합 트리 알고리즘에서 상태의 전이를 계산할 때 시간 t-1에서의 상태들로부터 시간 t의 상태로의 전이에서 모든 상태로부터의 전이를 고려하는 것이 아니라, 하나의 최대 전이값을 선택하여 계산한다. 이는 추론 과정에서 clique 노드 C_1에서 분리자 S_{12}로 메시지를 전달할 때, 분리자에 속하지 않는 노드들의 값들을 모두 더하는 대신 다음의 식과 같이 하나의 최대값만을 선택하여 메시지를 전달함으로써 계산할 수 있다.

$$f'(S_{12}) = \max_{C_1 \setminus S_{12}} f(C_1)$$

$f(C_1)$과 $f'(S_{12})$는 clique C_1의 potential 함수와 분리자 S_{12}의 갱신된 potential 함수를 각각 의미하며, $C1\backslash S_{12}$는 clique와 분리자에 속한 노드들의 차집합을 의미한다.

표 3은 '실행' 제스처에 대해 각 손의 움직임을 모델링하는 두 노드 X^1, X^2의 입력 방향 코드(O^1, O^2)에 대한 은닉 노드의 상태 변화를 디코딩한 결과이며, 그림 14는 이들 결과를 오른손과 왼손의 이동 궤적들에 매핑한 결과를 보여주고 있다. 그림 14에서 원 안의 숫자들을 입력 방향 코드 또는 방향 코드 공간을 분할하는 은닉 노드 X^1, X^2의 값을 의미한다. 노드들의 각 상태는 전체 제스처의 부분들을 표현하고 있다. 주어진 관측 데이타에 대해 각 손의 움직임을 모델링하는 은닉 노드들이 관측 데이타를 대체로 균일하게 분할하면서 직관적인 관점에서의 상태 전이를 하고 있는 것으로 보아 훈련된 모델이 제스처를 잘 모델링하고 있다고 할 수 있다.

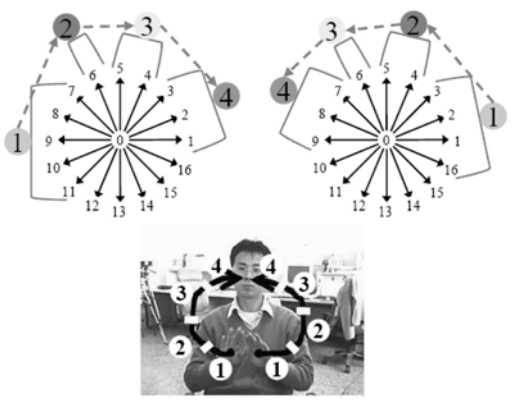


그림 14 '실행' 제스처에 대한 은닉 노드 X^1, X^2의 코드 분할 및 상태 전이

7. 결론 및 향후 과제

컴퓨터의 성능 및 정보 표현에 대한 기술 발달로 인해 컴퓨터와의 상호 작용을 위한 새로운 방법들이 요구되고 있다. 손은 신체의 다른 어떤 부분보다도 움직임이 자유로우며, 많은 것을 표현할 수 있으므로 컴퓨터와의 상호작용을 위해 손을 이용하는 것은 적절한 방법이 된다. 본 논문에서는 미디어 플레이어 또는 PowerPoint™ 제어를 위한 10가지의 손 제스처를 정의하고, 인식하는 방법을 제안하였다.

확률 모델이 잡음이나 불완전한 관측값을 표현하는데 유용하긴 하지만 성공적인 인식을 위해서는 양질의 입력값을 가지는 것이 필요하다. 영상 처리 단계에서 손의 정확한 검출을 위해 Haar-like 얼굴 검출기[22]를 통해 추출된 얼굴 영역의 픽셀들로부터 생성된 피부색 모델과 YIQ 색상 모델을 결합한 방법을 적용하였다. 손의 비선형적 또는 비연속적인 움직임에서도 강인한 추적을 위해 Argyros 등[20]의 방법을 기반으로 광류를 적용한 새로운 방법을 제안하였다.

DBN 프레임워크를 이용하여 손 제스처 인식 모델을 제안하였으며, 제안하는 모델은 손의 움직임을 표현하는 방향 코드, 두 손의 상대적 위치 및 얼굴과 손의 상대적 위치를 관측값으로 가진다. 제안한 모델의 구조는 coupled HMM에서의 완전히 결합된 형태를 완화시키고, coupled HMM에서는 표현하지 못하는 두 확률 과정의 공유 정보를 표현하였다. 제안한 DBN에 대한 독립 제스처 인식 성능은 490개의 실험 비디오 데이타로 7-fold 교차 검증 방법으로 평가하여 평균 99.59%의 인식률을 얻었다.

본 논문에서는 인식에 필요한 모든 특징들을 양자화하여 사용하였다. 이는 양자화 과정에서 보다 좋은 성능을 위해 필요한 일부의 정보를 잃어버렸을 수도 있다. 그러나, 이 논문에서 제안 방법은 수화 인식이나 전신 제스처 인식과 같은 보다 복잡한 제스처 인식에 대한 향후 연구로의 이정표가 될 수 있다.

참고문헌

[1] G. Johansson, "Visual Perception of Biological

표 3 관측 데이타 O^1, O^3에 대한 은닉 노드 X^1, X^2의 상태 전이

O1	11	10	10	9	0	8	8	7	7	6	6	6	0	6	6	6	5	6	5	5	5	4	0	4	3	3	3	2	2	3	2	2	2	2	1	0
O3	16	16	1	1	0	2	2	3	3	3	4	4	0	4	4	5	5	5	5	6	6	6	0	6	7	7	7	7	7	8	8	8	9	8	8	8
X1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4
X2	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4

Motion and a Model for Its Analysis," *Perception and Psychophysics*, Vol.14, pp. 201-211, 1973.

[2] J. Arggarwal and Q. Cai, "Human Motion Analysis - A Review," *Computer Vision and Image Understanding*, Vol.73, pp. 428-440, 1999.

[3] V. Pavlovic, R. Sharma, and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp. 677-695, 1997.

[4] C. Myers and L. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition," *The Bell System Technical Journal*, Vol.60, pp. 1389-1409, 1981.

[5] J. Yamato, J. Ohya, and K. Ishii, "Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model," In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Champaign, USA, pp. 379-385, June 1992.

[6] H.-I. Suk and B.-K. Sin, "HMM-Based Gait Recognition with Human Profiles," In *Proceedings of Joint IAPR International Workshops SSPR 2006 and SPR2006*, Hong Kong, China, pp. 596-603, August 2006.

[7] H.-K. Lee and J.-H. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.21, No.10, pp. 961-973, 1999.

[8] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 994-999, June 1997.

[9] F. Jensen, *Bayesian Networks and Decision Graphs*, Chapter 1, pp. 3-34, Springer, 2001.

[10] Y. Du, F. Chen, W. Xu, and Y. Li, "Recognizing Interaction Activities using Dynamic Bayesian Network," In *Proceedings of IEEE International Conference on Pattern Recognition*, Hong Kong, China, Vol.1, pp. 618-621, August 2006.

[11] S.-H Park and J. Aggarwal, "A Hierarchical Bayesian Network for Event Recognition of Human Actions and Interactions," *ACM Journal of Multimedia Systems*, Vol.10, No.2, pp. 164-179, 2004.

[12] H. Avilés-Arriaga, L. Sucar, and C. Mendoza, "Visual Recognition of Similar Gestures," In *Proceedings of IEEE International Conference on Pattern Recognition*, Hong Kong, China, Vol.1, pp. 1100-1103, August 2006.

[13] V. Pavlovic, *Dynamic Bayesian Networks for Information Fusion with Applications to Human-Computer Interfaces*, Ph. D. Dissertation, University of Illinois at Urbana-Champaign, 1999.

[14] A. Wilson, *Adaptive Models for the Recognition of Human Gestures*, Ph. D, Dissertation, MIT Program in Arts and Sciences, 2000.

[15] N. Oliver and E. Horvitz, "A Comparison of HMMs and Dynamic Bayesian Networks for Recognizing Office Activities," *User Modeling*, pp. 199-209, 2005.

[16] R. León, "Continuous Activity Recognition with Missing Data," In *Proceedings of IEEE International Conference on Pattern Recognition*, Quebec, Canada, Vol.1, pp. 439-446, August 2002.

[17] M. Yang and N. Ahuja, "Recognizing Hand Gestures Using Motion Trajectories," In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Fort Collins, USA, Vol.1, pp. 23-25, June 1999.

[18] A. Nefina, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *Journal of Applied Signal Processing*, Vol.11, No.1, pp. 1-15, 2002.

[19] S. Wong and R. Cipolla, "Continuous Gesture Recognition Using a Sparse Bayesian Classifier," In *Proceedings of IEEE International Conference on Pattern Recognition*, Hong Kong, China, Vol.1, pp. 1084-1087, August 2006.

[20] A. Argyros and M. Lourakis, "Real-Time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera," In *Proceedings of European Conference on Computer Vision*, Prague, Czech Republic, Vol.3, pp. 368-379, May 2004.

[21] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A Survey on Pixel-Based Skin Color Detection Techniques," *Pattern Recognition*, Vol.40, No.3, pp. 1106-1122, 2007.

[22] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, Vol.57, No.2, pp. 137-154, 2004.

[23] G. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface," Intel Technology Journal Q2, pp. 1-15., 1998.

[24] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol.77, pp. 257-285, 1989.

[25] C. Bishop, *Pattern Recognition and Machine Learning*, Chapter 8, pp. 359-422, Springer, 2007.

[26] K. Murphy, *Dynamic Bayesian Network: Representation, Inference and Learning*, Ph.D. Dissertation, University of California, Berkeley, 2002.

[27] C. Huang and A. Darwiche, "Inference in Belief Networks: A Procedural Guide," *International Journal of Approximate Reasoning*, Vol.15, No.3, pp. 225-263, 1994.

[28] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*,

Vol.39, No.1, pp. 1-38, 1977.
[29] http://bnt.sourceforge.net/
[30] http://sourceforge.net/projects/opencvlibrary/

석 홍 일

2004년 부경대학교 멀티미디어공학과 졸업(학사). 2007년 부경대학교 컴퓨터공학과 졸업(석사). 2008년 고려대학교 컴퓨터학과 박사과정. 관심분야는 패턴인식, 컴퓨터 비전, 인공지능 등

신 봉 기

1985년 서울대학교 자원공학 학사. 1987년 한국과학기술원 전산학 석사. 1995년 한국과학기술원 전산학 박사. 1987년~1999년 한국통신 멀티미디어연구소. 1999년~현재 부경대학교 전자컴퓨터정보통신공학부 부교수. 관심분야는 인공지능, 패턴인식, 컴퓨터시각, 기계학습 등