비전 기반 휴먼 제스처 자동 분석 기술

노명철* · 양희덕** · 이성환***

1. 서 론

최근 들어 지능형 환경, 휴머노이드 로봇 등을 비롯하여 사람의 행동을 자동으로 분석하고, 능동적인 서비스를 제공하는 서비스 기술이 각광을 받고 있다. 이러한 자동화된 서비스를 제공하기 위하여서는 휴먼-컴퓨터 상호 작용 기술이 필수적이고, 별도의 부가장치가 필요 없는 비전 기반 인터페이스는 자연스러운 휴먼-컴퓨터 상호작용을 위한 대표적인 인터페이스이다. 비전 기반으로 제스처를 분석하여 상호작용 함으로써, 자연스럽고, 편리하며, 진보된 인터페이스를 제공할 수 있다.

비전 기반의 제스처 연구는 목적에 따라서 몇 가지로 구분될 수 있다. 사용되는 데이터의 범위에 따라서 수화 분석을 위한 손 제스처 연구, 휴먼 몸체의 추적 및 포즈 재구성을 위한 상체 제스처 연구, 걸음걸이 인식을 위한 전신 제스처로 제스처 연구를 나눌 수 있다. 휴먼-컴퓨터 상호작용을 위한 제스처 인식 기술을 목적에 따라서 분류하면 지시형 제스처와 대화형 제스처 기술로 나눌 수 있다. 지시형 제스처는 손을 이용하여 컴퓨터 혹은 로봇에게 특정 방향 또는 물건 등을 지시하는 제스처이고, 대화형 제스처는 휴먼-로봇 간의 간단한 대화, 명령을 전달 할 수 있는 제스처로, 수화, 지화, 명령형 제스처로 다시 나눌 수 있다. 명령형 제스처는 손을 이용하여 일상생활에서 사용되는, '가라', '와라', '오른쪽으로 가라', '정지' 등의 단순한 명령을 나타내는 제스처를 나타낸다. 또한, 방법론에 따라서 3차원 모델링 기반 제스처 인식과 2차원 템플릿기반 제스처 인식으로 나눌 수 있다.

본 논문에서는 이러한 휴먼-컴퓨터 상호작용을 위한 3차원 인체 자세 재구성 기술, 대화형 제스처 인식 기술, 시점 변화에 무관한 템플릿 기반 제스처 인식 기술과 제스처 인식 알고리즘을 객관적으로 성능 평가할 수 있는 데이터베이스를 소개한다.

2. 3차원 인체 자세 재구성 기술[1]

시점에 강인한 제스처를 분석하기 위하여서는 입력 비디오 영상으로부터 사람 영역을 추출하고 3차원으로 인체를 모델링하는 기술이 필수적이

※ 교신저자(Corresponding Author) : 이성환, 주소 : 서울특별시 성북구 안암동 5가 1번지 고려대학교 정보통신대학 컴퓨터 · 통신공학부(136-713), 전화 : 02)3290-3572, FAX : 02)3290-4280, E-mail : swlee@image.korea.ac.kr
* 고려대학교 컴퓨터 · 통신공학부 박사과정 (E-mail : mcroh@image.korea.ac.kr)
** 고려대학교 컴퓨터 · 통신공학부 박사과정 (E-mail : hdyang@image.korea.ac.kr)
*** 고려대학교 컴퓨터 · 통신공학부 정교수

다. 본 절에서는 스테레오 카메라를 이용한 3차원 인체 자세 재구성 기술을 소개한다.

2.1 3차원 인체 모델

3차원 인체 모델링은 인체 구성요소의 모양 및 움직임을 제어한다. 인체의 모델은 관절과 마디로 구성된 3차원 구조를 가지고 있고, 17개의 마디로 구성되어 있으며, 37개의 자유도를 갖고 있다. 또한, 손, 발, 머리에 관절의 끝을 알려주는 표시자를 두었다. 이 표시자는 예측된 관절의 위치 정보로부터 각도 정보를 추출할 때 사용된다. 그림 1은 제안된 3차원 인체 모델의 형태와 계층적 구조를 보여주고 있다. 각 인체 구성요소들은 실제의 인체 모양과 비슷한 모양을 갖도록 Superquadrics를 사용하여 표현하였다.

인체 구성요소간의 움직임을 위해서 전방향/역방향 운동학을 생성하였다. 인체 구성요소의 상위 마디의 움직임은 마디와 연결되어 있는 모든 하위 마디의 움직임에 영향을 준다. 이를 Kinematics chain이라 한다. 예를 들면, 그림 1에서 Upper Torso, Neck, Head가 하나의 Kinematics Chain이다. 또한, 실 환경의 원근을 표시할 수 있는 Perspective Camera Model을 이용하여 깊이 영상과 실루엣 영상을 생성한다. 그림 2는 3차원 인체 모델로부터 생성된 깊이 영상과 실루엣 영상을 보여주고 있다.

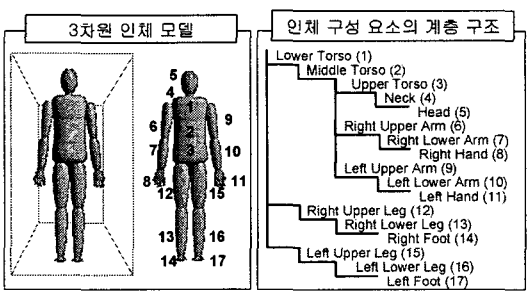


그림 1. 3차원 인체 모델

2.2 3차원 자세 재구성

사람의 자세를 2차원의 깊이 영상과 이에 대응되는 3차원 좌표 정보로 표현할 수 있다. 입력된 2차원 깊이 영상이 다양한 자세의 2차원 깊이 영상들의 선형 결합으로 표현 가능하다면, 다양한 자세들에 대응되는 3차원 좌표 정보의 선형 결합으로 입력된 2차원 깊이 영상에 대응되는 3차원 모델을 재구성할 수 있다. 그림 3은 프로토타입 영상들의 선형 결합으로 3차원 인체 재구성하는 예를 보여주고 있다.

깊이 영상 $s=(s'_1,\cdots,s'_n)^T$는 0~255의 값을 갖는 벡터로 나타내고, 3차원 인체 모델은 $p=((x_1,y_1,z_1),\cdots,(x_q,y_q,z_q))^T$로 나타내어지고, n은 픽셀의 개수, x,y,z는 인체 모델 관절의 3차원 좌표를 나타내고, q는 관절의 개수를 나타낸다. 전체 3차원 데이터는 다음 식(1)과 같이 나타내어지며, m은 전체 프로토타입의 개수를 나타낸다.

그림 2. 3차원 인체 모델의 깊이와 실루엣 영상

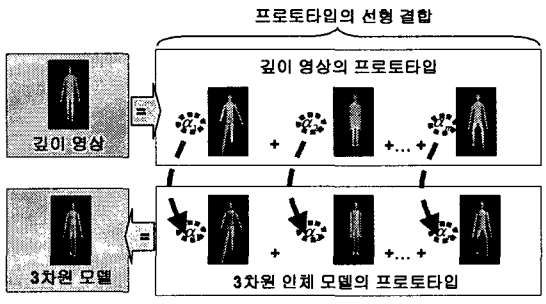


그림 3. 3차원 자세 재구성

$$S=(s_1,\cdots,s_m),\ P=(p_1,\cdots,p_m) \qquad (1)$$

2.3 계층 구조

다양한 형태의 인체 자세의 모델을 학습 데이터로 사용하기 위해서 3차원 인체 모델을 이용하여100,000 개 이상의 데이터를 생성하였다. 생성된 데이터는 사람이 표현 할 수 있는 많은 형태의 자세를 갖고 있으며, 각 인체 구성 요소 별로 움직일 수 있는 범위에 대한 제약을 적용하여 생성되었다. 많은 양의 데이터로 인하여 검색 속도가 느려지는 단점을 해결하기 위해서, 학습 데이터를 3단계의 계층 구조를 가지는 형태로 분류하였다. 학습 데이터를 분류하기 위해서 K-means 군집화 알고리즘을 사용하였다. 클러스터의 1, 2 단계 분류를 위해서는 실루엣 영상을 이용하였고, 3 단계에서는 깊이 영상을 이용하였다. 1, 2 단계에서는 자세의 모양을 분류하였고, 3단계에서는 비슷한 자세를 갖는 데이터의 전/후 관계를 분류하였다. 각 클러스터들은 2차원 실루엣 영상 공간에서 유사한 모양을 갖는다. 계층적 분류를 위해서 하위 단계는 상위 단계의 클러스터의 평균값을 갖도록 구성한다.

2.4 시-공간 특징 정보

한 장의 실루엣 영상을 이용하면 현재 영상의 노이즈에 민감하기 때문에 재구성 단계의 상위 레벨에서는 실루엣 영상의 누적 영상, H(x,y,t)를 이용하였다. 실루엣 영상의 누적 영상을 이용함으로써, 현재 자세와 비슷한 자세를 구성할 수 있다.

$$H_t(x,y,t)=\begin{cases}\tau & ,if\ D(x,y,t)=1\\ max\,(0,H_t(x,y,t-1)-\lambda) & ,otherwise\end{cases} \qquad (2)$$

3. 대화형 제스처 분석 기술

지능형 휴머노이드 로봇을 위하여서는 사람과 자연스러운 상호 작용 기술의 개발이 필수적이다. 손동작과 표정 등의 비수지적 표현을 이용하여 대화의 내포적인 의미 및 감성 인식도 능동적인 서비스 제공을 위하여 필요하므로 수화, 지화의 연구와 더불어 많이 연구되고 있다[2]. 이러한 기술들 중에서 기타 부착 장치 없이, 가장 자연스럽게 상호 작용할 수 있는 방법은 비전을 기반으로 둔 손 제스처 인식 기술이다. 사람의 제스처 중에서 가장 중요하고, 많은 의미를 가지고 있는 것은 손과 팔을 이용하는 손 제스처라고 할 수 있다. 대표적인 예로 수화, 지화, 명령형 제스처를 들 수 있다. 사람의 움직임 변화에 강인한 인식을 위하여서는 가려짐과 방향을 고려한 3차원 손 제스처 인식이 필수적이다.

본 절에서는 수화, 지화의 기본 기술인 3차원 손 포즈 추정과 3차원 손 제스처 인식을 이용한 제스처 인식을 소개한다.

3차원 손 제스처는 움직임 변화, 회전 등에 강

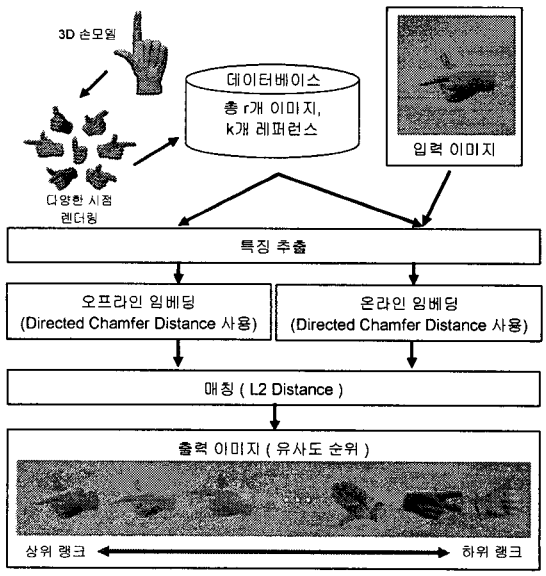


그림 4. 3차원 손 포즈 인식을 위한 시스템 구성

인한 손 모양 인식과 손, 팔의 움직임을 분석하는 두 단계가 필요하다. 또한, 정확한 손, 팔의 움직임을 추적하기 위하여서 입력 비디오 영상으로부터 3차원 휴먼 모델로의 재구성 방법이 필요하다.

3차원 손 포즈 추정을 이용하여 각도 변화와 자기 겹침(Self Occlusion)에 강인한 손, 팔 추적을 할 수 있고, 이러한 움직임을 분석함으로써 명령형 제스처를 인식한다. 수화 인식은 이러한 손, 팔의 움직임과 손 모양 인식을 기본으로 한다. 또한, 수화 인식은 얼굴, 어깨 등의 손이 아닌 다른 신체 부위의 움직임을 통해 나타내어지는 비수지(non-manual) 표현을 분석이 필요하다.

본 논문에서는 지화 및 수화 인식으로 확장할 수 있는 3차원 손 모양 추정과 3차원 손/팔의 움직임을 이용한 명령형 제스처 인식 기술을 다룬다.

3.1 명령형 제스처 인식

명령형 제스처는 손의 추출 및 추적을 통해 얻어진 궤적으로부터 특징을 추출하고, 특징 변화를 동적 프로그래밍(Dynamic Programming)을 이용하여 인식된다. 특징으로는 입력 동영상으로부터 추적되는 손의 위치, 속도, 각도를 사용한다. 다음 수식 (3), (4), (5) 은 각각의 특징을 나타낸다.

$$l_t = \frac{L_t}{L_{\max}}, \quad where\, L_{\max} = \max_{t-1}^{n}(L_t) \tag{3}$$

$$\theta_1 = \alpha\tan2(d_{y1}, d_{x1}),\ \theta_2 = \alpha\tan2(d_{y2}, d_{x2})$$
$$where\ d_{x1} = X_t - C_x,\ d_{y1} = Y_t - C_y,\quad d_{x2} = X_t - X_{t+1},\ d_{y2} = Y_t - C_{t+1} \tag{4}$$

$$v_t = \frac{V_t}{V_{\max}}, \quad where\, V_{\max} = \max_{t-1}^{n}(V_t) \tag{5}$$

(C_x, C_y)은 궤적에서의 무게중심, l_t는 시간 t에서 중심과의 거리, Θ_1는 무게중심과 현재 점과의 각도, Θ_2는 연속된 두 점사이의 각도, v_t는 연속된 두 점 사이의 속도를 나타낸다. L_{max}는 중심점과 어떠한 점 사이로부터 가장 긴 거리, V_{max}는 두 점 사이의 최대 속도 값을 나타낸다. l_t와 V_{max}는 정규화를 통하여 0~1 사이의 값을 가지도록 한다.

제스처 인식을 위하여서 모델 제스처의 특징 템플릿과 입력 제스처의 특징 템플릿 사이의 최소 거리를 동적 프로그래밍을 이용하여 계산한다[3]. 입력 제스처의 특징 템플릿과 가장 작은 거리를 가지는 데이터베이스에 있는 모델 제스처의 특징 템플릿을 찾음으로써 수행된다.

3.2 3차 손 포즈 추정

3차원 손 포즈 추정은 데이터베이스를 생성하는 과정, 오프라인과 온라인 임베딩 프로세스의 세 단계로 이루어진다. 그림 4는 포즈 추정을 위한 시스템을 보여준다.

손은 16개의 링크로 모델링 된다. 1개의 손바닥과 5개의 손가락으로 각각의 손가락은 3개의 링크로 구성되어 있으며 20 DOF(degree of freedom)를 갖는다. 이러한 형태 변수와 시점 변수를 더하여 총 23가지의 DOF를 갖는다. 손의 형태를 나타내는 벡터 C_h = (c_1, c_2,…,c_{20})와 시점을 나타내는 시점 매개변수 벡터 V_h = (v_1, v_2, v_3)가 주어질 때 한 장의 손 영상은 다음과 같이 23개의 포즈 파라미터 벡터로서 표현될 수 있다.

$$P_h = (\ c_1,\ c_2,\ c_3\ ,\cdots,\ c_{20},\ v_1,\ v_2,\ v_3\ ) \tag{6}$$

학습을 위한 손 영상은 컴퓨터그래픽 툴로 렌더링하여 얻어지며 이때 3차원 매개변수 정보들도 함께 데이터베이스에 저장된다.

오프라인 임베딩 프로세스는 데이터베이스의 r개의 이미지에 대한 각각의 k개의 레퍼런스 이미

지 사이의 거리를 임베딩 하여 그 결과를 저장하고, 온라인 임베딩 프로세스에서는 입력이미지에 대한 k개의 레퍼런스 이미지 사이의 거리를 임베딩 하여 이 값을 오프라인으로 미리 계산되어 저장된 값과 매칭하여 가장 유사한 이미지가 선택되어진다.

모델의 포즈를 추정하기 위한 매칭은 Chamfer 거리를 Lipschitz 임베딩 방법을 이용해 근사화한 Approximated Directed Chamfer 거리를 이용한다[4]. Chamfer 거리는 에지 간의 거리를 측정하기에 효과적이며 노이즈에 강인한 널리 알려진 방법으로, 에지 이미지는 각각의 픽셀 위치에 대응하는 점들의 집합으로서 나타내어지며 임의의 영상 A에서 B로의 Directed Chamfer 거리, c(A, B)는 수식(7)로 정의된다. Undirected Chamfer 거리는 수식(8)과 같이 정의되며 이는 c(A, B)와 c(B, A)의 합으로 나타내어진다.

$$c(A,B) = \frac{1}{|A|}\sum_{a \in A} \min_{b \in B} | a - b | \qquad (7)$$

$$C(A,B) = c(A,B) + c(B,A) \qquad (8)$$

본 논문에서 Directed Chamfer 거리와 Undirected Chamfer 거리는 각각 DCD와 CD로 표기한다. 모델 이미지와 입력이미지간의 DCD는 모델의 에지 이미지가 n개의 점들로 구성되어있으며 데이터베이스에 총 d개의 이미지가 있을 때, 시간 복잡도는 $O(n\log n)$이다. 따라서 데이터베이스의 이미지개수인 d와 에지를 구성하는 픽셀수 n이 커짐에 따라 시간이 많이 걸리며 많은 메모리를 필요로 하게 된다.

다차원공간상의 거리를 저차원 공간상으로 임베딩함에 있어서, 임의의 공간 G에서 k차원 공간 R_k상으로 임베딩하였을 때 공간 G상의 두 점간의 거리가 공간 R_k상에서 왜곡을 최소화하면서 효과적으로 보존이 되는 것이 중요하다. 이러한 임베딩은 공간 G상에서의 거리측정방법의 계산량이 많은 경우 저차원 R_k상으로 매핑한 뒤 Lp norm의 연산으로서 복잡한 연산을 대체할 수 있으므로 유용하다. 본 논문에서는 Lipschitz embeddings를 사용하여 데이터베이스에 인덱싱하는 방식으로 계산복잡도 문제를 해결한다[2]. Lipschitz embeddings의 기본적인 아이디어는 두 개의 가까운 점은 제 삼의 점에 대하여 비슷한 거리를 갖는다는 것을 이용하는 것이다. 입력 에지이미지 g로부터 Rk상으로의 Lipschitz embedding E(g)는 아래의 식(9)와 같이 정의된다. 이 때 $r_1, r_2, \cdots, r_k$는 데이터베이스로부터 임의로 선택되어진 k개의 레퍼런스 이미지를 의미하며 c는 식(7)에서 정의되어진 DCD를 뜻한다. 이러한 방법을 Approximated Directed Chamfer 거리라한다.

$$E(g) = (c(g,r_1), c(g,r_2), ..., c(g,r_k)) \qquad (9)$$

4. 시점 변화에 무관한 제스처 인식 기술[5]

시각 기반의 제스처 인식 연구는 방법론에 따라 크게 다음 두 가지로 나뉠 수 있다. 첫 번째는 모델기반 방법으로, 입력 영상으로부터 2차원 또는 3차원 인체 모델을 이용하여 각 구성요소를 분석한다[6]. 입력 영상에서 인체 모델을 정합하고 역운동학 정보를 추출하여 팔과 다리를 찾아내는 과정에서 복잡도와 계산량이 높을 뿐 아니라, 오차가 누적되는 단점을 갖고 있다. 두 번째는 형상기반 방법이다[7]. 이 방법은 입력된 영상을 직접 분석하여 모션 정보를 추출하고 인식하는 방법으로, 알고리즘이 간단하여 실시간 처리가 가능하다는 장점이 있으나 시점에 종속적인 문제를 안고

있다. 이 절에서는 간단한 제스처 인식의 실시간 시스템을 위한 형상기반의 시점 무관 방법론을 소개한다.

4.1 공간 정규화

입력 영상에서 사람은 어디에서든 위치할 수 있으며, 카메라의 거리와 렌즈의 초점거리에 따라 실루엣 영상 및 모션 템플릿이 다른 크기와 다른 중심점을 가질 수 있다. 입력영상에서 배경모델을 통해 추출된 실루엣에서 주요영역을 구하고 그것의 높이를 기준으로 크기를 일정하게 했다.

4.2 시간 정규화

제스처의 속도는 다양한 사람뿐 아니라 같은 사람의 행동 시마다 편차를 갖는다. 이것은 형상기반 인식 시스템의 성능에 떨어뜨리는 주요한 원인 중의 하나이다. 이를 해결하기 위해 모션량을 정의하고 이를 이용하여 시간흐름에 따른 모션 히스토리 정보의 사라짐을 방지하여, 시간 정규화를 하였다. 연속되는 두 복원객체의 차이 D를 식 (10)과 같이 정의하고, 모션량 μ_t 를 식 (11)와 같이 계산한다. R_t, R_{t-1}은 t와 $t-1$시간의 3차원 객체를 나타낸다.

$$D_t(x,y,z) = |R_t(x,y,z) - R_{t-1}(x,y,z)| \quad (10)$$

$$\mu_t = \iiint D_t(x,y,z)dxdydz \quad (11)$$

모션량을 이용하여 모션이 일어나지 않은 프레임에 대하여는 이전 히스토리 정보를 보존하고, 행동이 발생한 양에 따라 이전 히스토리 정보를 사라지게 하여 이 문제를 그림 2. (c), (d) 같이 해결하였다.

그림 5. 시간 정규화 결과

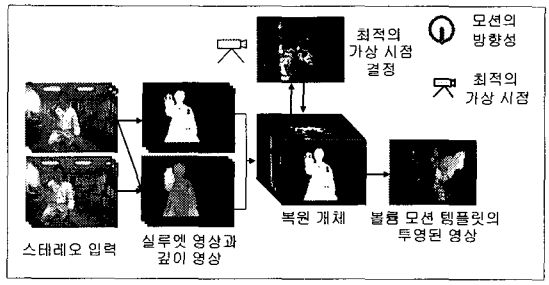


그림 6. 볼륨 모션 템플릿 생성을 위한 절차

4.3 Volume Motion Template(VMT)

볼륨 모션 템플릿 (VMT)은 깊이 정보를 통해 재구성된 3차원 템플릿이며, 모션 히스토리 정보를 3차원 공간상에 담고 있다. VMT는 기존 시점 종속적인 방법론의 문제점을 극복하기 위해 고안되었다. VMT 생성을 위한 절차는 다음과 같고, 그림 6은 VMT 생성 단계를 도식화하여 보여준다.

1) 배경 모델링을 통해 실루엣 영상을 구하고 스테레오 영상을 통하여 깊이 영상을 계산
2) 실루엣 영상과 깊이 영상을 통해 3차원 공간상에 복원 객체를 생성
3) 연속된 복원객체의 차이를 계산하고 모션량을 계산
4) 새로운 모션을 추가하고 이전 히스토리 정보를 모션량에 비례하게 감쇄시켜 VMT를 생성
5) VMT의 상단 투영 이미지에서 모션의 방향성을 이용하여 가상 시점 계산
6) 최적의 가상 시점에서 투영시켜 최종 모션 템플릿 생성

VMT는 3차원 공간에 표현되었으므로, 단순히 y축 회전만으로 시점을 바꿀 수 있다. 상단 시점에서 내려다 본 투영 이미지를 이용하여 최적의 가상 시점을 찾아 최적의 가상 시점에 대하여 수직인 방향으로 VMT를 투영시켜 최종 모션 템플릿을 생성한다.

5. KU 제스처 데이터베이스[8, 9]

제스처 알고리즘을 개발하고, 인식 실험을 하여 성능을 테스트하기 위하여서는 잘 디자인된 데이터베이스 구축이 필수적이다. KU 제스처 데이터베이스는 전신 제스처와 명령형 제스처의 두 개의 데이터베이스로 구성되어있고, 2차에 걸쳐서 데이터가 수집되었다. 첫 번째 기간에는 일상생활에서 일어날 수 있는 14개의 정상 제스처를 정의하여 20명의 60~80세의 노인들을 대상으로 전신 제스처 데이터를 수집하였고, 두 번째 기간에는 20명의 일반인을 대상으로, 위급한 상황에서 발생할 수 있는 10개의 비정상 제스처를 정의하여 전신 제스처 데이터를 수집하고, 30개의 명령형 제스처를 정의하여 전신과 상반신 제스처 데이터를 수집하였다.

일상생활의 제스처가 다양하지만, KU 제스처 데이터베이스에서는 노인들을 대상으로 (1)의자에 앉기, (2)의자에서 일어나기, (3)제자리 걷기, (4)무릎과 허리 짚기, (5)오른손 들기, (6)손 앞으로 뻗기, (7)허리 숙이기, (8)바닥에 앉기, (9)바닥에 무릎 꿇기, (10)바닥에 눕기, (11)손 흔들기, (12)제자리 뛰기, (13)앞으로 걷기, (14)원형으로 돌기의 14가지 제스처를 포함하고 있다. 비정상 제스처로는 바닥에 서있는 상태에서 (1)앞으로 쓰러지기, (2)뒤로 쓰러지기, (3)좌로 쓰러지기, (4)우로 쓰러지기와 바닥에 앉은 상태에서 (5)좌로 쓰러지기, (6)우로 쓰러지기, (7)뒤로 쓰러지기와 의자에 앉은 상태에서 (8)좌로 쓰러지기, (9)우로 쓰러지기, (10)앞으로 쓰러지기의 10가지 제스처를 포함하고 있다.

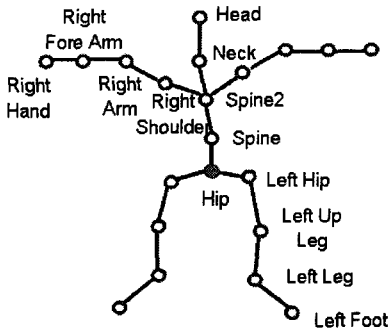


그림 7. KU 제스처 데이터 베이스의 신체 구성 요소의 계층적 구조

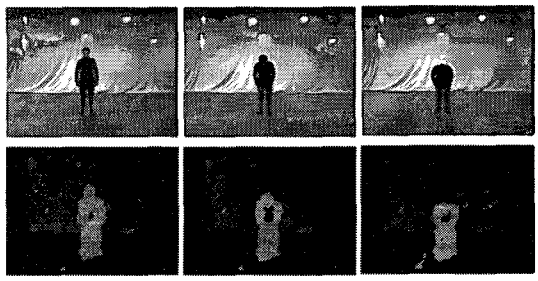


그림 8. 스테레오 입력 영상들 중 왼쪽 영상(위)과 깊이 영상(아래)

KU 제스처 데이터베이스는 3차원 제스처 데이터 파일, 세 가지 방향에서 촬영된 스테레오 비디오 데이터, 전경 영상 영역의 실루엣 비디오 데이터의 세 가지 종류 데이터를 포함하고 있다. 다양한 종류의 데이터를 제공함으로써 제스처 인식, 신체 관절 정보 추적, 3차원 좌표 생성 등의 다양한 연구 분야에 사용 될 수 있다. 그림 7은 신체 구성 요소의 3차원 좌표를 나타내기 위하여 사용된 계층적 구조를 보여주고, 그림 8은 스테레오 카메라로부터 입력된 한 쌍의 영상 중 왼쪽 영상

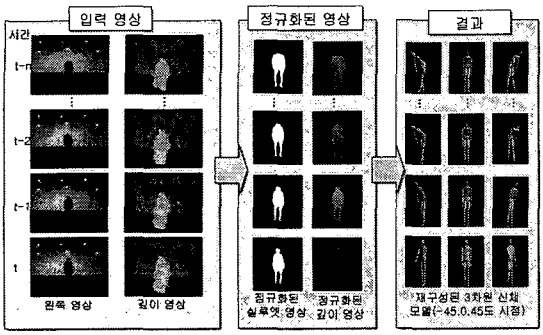


그림 9. 재구성된 3차원 인체 모델

들(위)과 한 쌍의 영상으로부터 계산되어진 깊이 영상의 예를 보여준다.

6. 실험 및 결과 분석

6.1 3차원 인체 모델 실험

성능평가를 위해서, KU 제스처 데이터베이스의 일부인 정상적 제스처 데이터를 이용하였다. 그림 9는 KU 제스처 데이터베이스의 인사하기 동작을 이용하여 실험한 결과를 보여준다. 입력 영상은 깊이 영상과 스테레오 영상이며, 결과 영상은 재구성된 3차원 인체 모델의 인체 구성 요소별 각도 정보이다.

실험 결과의 정확성을 검증하기 위해서 KU 제스처 데이터베이스의 Ground Truth와 예측된 3차원 인체 모델의 각 인체 구성 요소의 평균 오류율을 계산하였다. 실험 결과, 입력된 영상으로부터 재구성된 3차원 인체 모델의 자세가 입력 영상의 자세와 유사함을 볼 수 있다. 그림 10에서 Θ는 x, y 축에 투영된 각의 평균차를 Ψ는 y, z 축에 투영된 각의 평균차를 나타낸다. 실험 결과 움직임이 많은 다리 부분에서는 다른 부분보다 오류율이 높음을 볼 수가 있다.

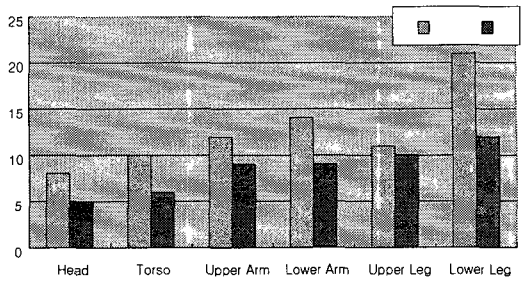


그림 10. 인체 구성 요소 별 평균 오류율

6.2 명령형 제스처 인식 실험

명령형 제스처 인식을 위한 제스처는 숫자 0~9까지를 쓰는 제스처 사용하여 실험을 하였다. 본 논문에서는 실험을 위하여 사용된 데이터는 두 가지 종류로, KU 제스처 데이터베이스의 숫자 쓰기 데이터 총 100개와 웹 카메라를 이용하여 촬영한 데이터 50개로 총 150개이다. KU 제스처 데이터베이스를 이용한 실험은 신체 구성 요소 정보 파일을 이용하여 손의 좌표를 추출하였다. 그림 11은 입력된 제스처와 데이터베이스에 있는 모델 제스처와의 거리를 측정하는 동적 프로그래밍 테이블의 예를 보여준다.

표 1은 각 실험 데이터에 따른 인식결과를 보여준다. KUGDB 데이터는 KU 제스처 데이터베이스를 사용한 결과를, 웹 카메라는 웹 카메라로 촬영한 영상을 사용한 결과를 각각 보여준다. 여러 가지 잡영과 추적의 오류로 인해서 웹 카메라를 사용하여 실험한 결과가 KUGDB 데이터를 사용한 결과에 비해 14%정도 저하되는 결과를 보여준다.

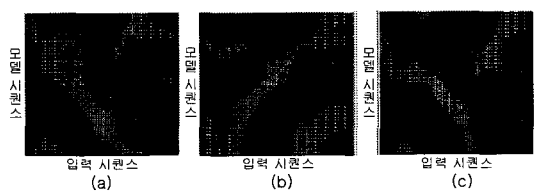


그림 11. 입력 제스처('5')와 모델('5', '6', '7')에 대한 동적 프로그래밍 테이블

표 1. 명령형 제스처 인식 실험 결과

결과 \ 동영상	KUGDB 데이터	웹 카메라
Detection rate(%)	88%	74%
False matches	12/100	13/50

오류는 서로 다른 숫자이지만, '1' 과 '9', '7' 등과 같이 특정 숫자의 궤적이 다른 숫자의 궤적에 포함되는 경우와 '6'과 '0'과 같이 비슷한 궤적을 보이는 제스처에서 발생하였다.

6.3 3차원 손 포즈 추정 실험

테스트를 위해서 그림 12와 같이 4가지 프로토타입을 정의하고 데이터 셋을 구성하였다. 데이터베이스 구성은 각 프로토타입에 대하여 수직 ±22.5도, 수평 ±90도 이내의 범위에서 30가지 시점으로부터 본 손 모양을 렌더링하여 사용하였다. 총 120장의 이미지 중 100장의 이미지는 오프라인 학습을 위해 사용하였고 나머지 20장의 영상은 테스트에서의 입력 영상으로 사용하였다. 데이터 셋의 100장의 이미지 중에서 20%는 레퍼런스 이미지로 (k=20) 사용되었다.

테스트에서 입력 이미지는 임의의 시점을 갖는 손 이미지이며 위치와 크기는 고정되었다고 가정하였다. 출력 이미지는 입력 이미지와 가장 유사한 이미지로서 매칭 결과가 최소인 이미지부터 차례로 10개의 이미지를 순위를 매겨 출력하였다.

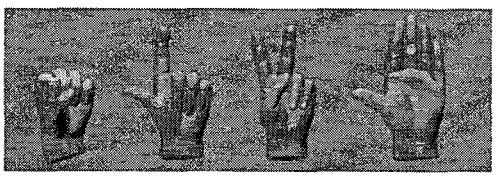


그림 12. 4가지 손 모양의 프로토타입

표 2. 손 포즈 추정 실험 결과

방법	1	1 - 3	1 - 6	1 - 12	Median
F, ADCD	85 %	90 %	100 %	100 %	1
F, DCD	75 %	90 %	95 %	100 %	1
C, ADCD	25 %	75 %	75 %	100 %	3
C, DCD	30 %	40 %	55 %	65 %	5

표 2는 실험 결과를 보여준다. 'F'는 잡영 없는 깨끗한 영상에 대하여 테스트한 것이고 'C'는 임의의 잡영을 생성하여 추가한 영상에서 테스트한 것을 의미한다. ADCD는 Approximated DCD 방법으로 테스트 한 것이고 DCD는 Directed Chamfer 거리를 사용한 결과를 보여준다. Median은 최상위 랭크들의 중간 값을 나타낸다. 실험결과 깨끗한 에지 이미지들을 사용해 실험한 결과 이 논문의 ADCD의 방법을 사용한 결과가 DCD만을 사용한 결과보다 인식율이 높게 나타났다. 잡영이 심하게 있는 이미지를 이용하여 테스트 결과 인식율은 깨끗한 에지이미지를 사용한 것 보다 떨어졌지만 DCD에 비해 ADCD를 사용한 쪽의 인식율이 상승한 것을 볼 수 있다.

6.4 시점 변화에 강인한 제스처 인식 실험

VMT의 효용성을 테스트하기 위해 0° ~ 90°사이의 7가지 다른 시점에서 획득한 제스처 영상에 대하여 실험하였다. 그림 13은 시점 변화에 따른 VMT와 MHI의 각각의 유사도를 보여준다. MHI는 시점이 변함에 따라 모션정보를 담고 있는 템플릿이 다양하게 변하며, 유사도 또한 떨어지는 면, VMT는 시점이 다양하게 변하여도 어느 정도 안정적인 유사도를 보여준다. 이는 시점에 독립적인 제스처 인식을 할 수 있음을 의미한다.

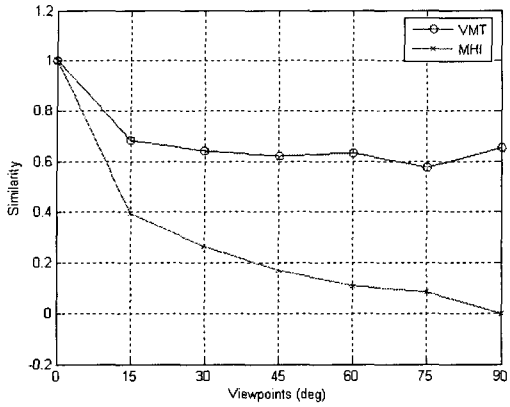


그림 13. 시점 변화에 따른 VMT와 MHI 템플릿 유사도

7. 결 론

차세대 컴퓨터 비전의 응용을 위해서는 사람의 움직임을 통하여 의도를 파악하고, 지능화된 서비스를 제공할 수 있는 휴먼-컴퓨터 인터페이스 기술의 개발이 필수적이다.

본 논문에서는 휴먼-컴퓨터 인터페이스를 위한 깊이 영상을 이용한 이에 대응되는 3차원 인체 모델의 선형 결합에 기반한 3차원 인체 자세 재구성 방법, 대화형 제스처 인식을 위한 3차원 손 포즈 인식 방법과 손 궤적을 이용한 명령형 제스처 인식 방법, 실시간 제스처 인식을 위한 시점 무관한 VMT 방법, 그리고 제스처 인식 실험을 위한 KU 제스처 데이터베이스를 소개하였다.

실생활에 사용될 수 있는 제스처 인식을 위하여서는 본 논문에서 다룬 인체 재구성, 명령형 제스처 인식 기술, 3차원 손 포즈 인식 기술이 기본되어져야하고, 본문에서 각 기술별로 언급하고 있는 바와 같이 3차원 공간과 실시간 처리 속도의 고려가 필요하다. 이러한 제스처 분석 기술은 로봇 제어, 노인 복지를 위한 서비스, 게임, 수화/지화 통역, 무인 감시 시스템 등을 비롯한 다양한 분야에서 사용될 수 있고, 이미 많은 활용 분야에 대한 활발한 연구가 수행중이다.

참 고 문 헌

[1] H.-D. Yang, A.-Y. Park, and S.-W. Lee, "Gesture Spotting and Recognition for Human-Robot Interaction," IEEE Trans. on Robotics, Vol. 22, 2006. (To appear)

[2] W. Gao, J. Ma, S. Shan, X. Chen, W. Zheng, H. Zhang, J. Yan, and J. Wu, "HandTalker: A Multimodal Dialog System Using Sign Language and 3-D Virtual Human," Proc. International Conference on Multimodal Interfaces, Beijing, China, pp. 564-571, October 2000.

[3] J. Alon, V. Athitsos, and S. Sclaroff, "Accurate and Efficient Gesture Spotting via Pruning and Subgesture Reasoning," Computer Vision in Human-Computer Interaction, Lecture Notes in Computer Science, Vol. 3766, pp. 189-198, 2005.

[4] V. Athitsos and S. Sclaroff, "Estimating 3D Hand Pose from a Cluttered Image," Proc. IEEE Conference on Computer Vision and Pattern Recognition, Wisconsin, pp. 432-439, June 2003.

[5] M.-C. Roh, H.-K. Shin, S.-W. Lee and S.-W. Lee, "Volume Motion Template for View-invariant Gesture Recognition," Proc. 18th IAPR/IEEE International Conference on Pattern Recognition, Hong Kong, Vol. 2, August 2006, pp. 1229-1232.

[6] C. Sminchisescu and B. Triggs, "Estimating Articulated Human Motion With Covariance Scaled Sampling," International Journal of Robotics Research, Vol. 22, No. 6, pp. 371-391, 2003.

[7] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," IEEE Trans. on Pattern Analysis

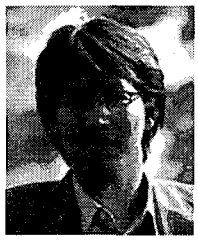
and Machine Intelligence, Vol. 23, No. 7, pp. 257-267, 2001.

[8] B.-W. Hwang, S. Kim and S.-W. Lee, "A Full-Body Gesture Database for Automatic Gesture Recognition," Proc. 7th IEEE International Conference on Automatic Face and Gesture Recognition, Southampton, United Kingdom, pp. 243-248, 2006.

[9] http://gesturedb.korea.ac.kr

이 성 환

- 1984년 서울대학교 전산학(학사)
- 1986년 한국과학기술원 전산학(석사)
- 1989년 한국과학기술원 전산학(박사)
- 1989년~1995년 충북대학교 컴퓨터과학과 조교수
- 1995년~2001년 고려대학교 컴퓨터학과 부교수
- 1997년~현재 고려대학교 인공시각연구센터 소장
- 2001년~2002년 MIT AI Lab. 객원교수
- 2001년~현재 고려대학교 컴퓨터·통신공학부 정교수
- 관심분야 : 컴퓨터 비전, 패턴인식, 영상처리



노 명 철

- 2001년 강원대학교 전산학(학사)
- 2003년 고려대학교 전산학(석사)
- 2003년~현재 고려대학교 전산학(박사과정)
- 관심분야 : 행동 분석, 제스처 인식, 로봇 비젼

양 희 덕

- 1999년 충남대학교 전산학(학사)
- 2003년 고려대학교 전산학(석사)
- 2003년~현재 고려대학교 전산학(박사과정)
- 관심분야 : 제스처 인식, 로봇 비젼, 얼굴 인식