시점에 독립적인 제스처 인식을 위한 볼륨 모션 템플릿

신호근, 이성환
고려대학교 정보통신대학 인공시각연구센터 / 컴퓨터학과
{hkshin, swlee}@image.korea.ac.kr

Volume Motion Template For View Independent Gesture Recognition

H.-K. Shin and S.-W. Lee
Dept. of Computer Science and Engineering, Korea University

요 약

본 논문은 시점에 독립적인 제스처 인식을 위하여 볼륨 모션 템플릿을 제안한다. 기존 제스처 연구에서 시점 문제와 행동 속도의 편차는 중요하면서도 어려운 문제이다. 첫째, 시점 문제는 하나의 단안 카메라나 스테레오 카메라를 이용하는 단방향 카메라 환경에서 발생하며 해결하기 어려운 문제이다. 모든 시점에서 학습시켜야 하는 기존 연구의 단점을 해결하기 위해, 다양한 시점입력에 독립적으로 인식을 할 수 있는 볼륨 모션 템플릿을 제안한다. 볼륨 모션 템플릿은 깊이 정보와 모션의 방향성 통해 최적의 가상 시점을 제공한다. 또한 볼륨 모션 템플릿을 이용하여 시스템의 신뢰성과 확장성 또한 개선하였다. 두 번째, 제스처가 발생 시마다 생기는 속도의 편차 문제이다. 입력 제스처의 시간-정규화를 통해 해결할 수 있는데, 시간 정보 대신 모션 량을 사용하여 이를 해결하였다. 볼륨 모션 템플릿을 이용하여 다양한 시점 입력에 대해 실험하였고, 기존 모션 히스토리 이미지와 비교하여 시점에 독립적인 결과를 얻었다.

1. 서 론

휴머노이드 로봇의 빠른 발전은 우리 생활을 로봇이 보조해줄 날이 머지않았음을 시사한다. 이에 따라 로봇의 형태뿐 아니라 인간과 친숙한 상호 인터페이스 또한 매우 중요한 기술로 부각되고 있다. 비전 센서 기반의 제스처 인식은 컴퓨터와 인간간의 진보적인 인터페이스 영역에서도 중요한 위치를 차지하고 있으며, 이러한 연구가 활발히 진행되고 있다.

이러한 비전 센서를 통한 제스처 인식 연구는 크게 다음 두 가지로 나뉠 수 있다. 첫 번째는 모델기반 방법이다[2]. 이 방법은 입력 영상으로부터 2차원 또는 3차원 인체 모델을 이용하여 각 구성요소를 분석한다. 이 연구에서는 입력 영상으로부터 인체의 관절정보를 추출하는 작업이 핵심이라 하겠다. 하지만 입력 영상에서 인체 모델을 정합하고 역운동학 정보를 추출하여 팔과 다리를 찾아내는 과정에서 복잡도와 계산량이 높을 뿐 아니라, 오차가 누적되는 단점을 갖고 있다. 특히 다수의 카메라 환경이 아닌 경우 신뢰성이 매우 떨어진다. 두 번째는 형상기반 방법이다[1, 3, 4, 5]. 이 방법은 입력된 영상을 직접 분석하여 모션 정보를 추출하고 인식하는 방법이다. 이 방법은 비교적 하위 단계에서 인식하는 시스템으로 알고리즘이 간단하고 가벼워 실시간 처리가 가능하다는 장점이 있다. 대표적인 예로는 모션 히스토리 이미지 (MHI)[1]가 있으나, 시점에 종속적인 문제를 안고 있다.

2. 기존 연구

서론에서 언급했듯이 모션 히스토리 이미지는 간단하며, 직관적인 알고리즘 중의 하나이다. 하지만, 2차원상의 움직임 정보만 이용하기 때문에 다음과 같은 단점이 있다.

첫 번째, 시점 기반 방법들은 학습에 필요한 각 시점별 형판을 모두 요구한다. 만일 시스템이 다양한 시점에서의 입력에서 인식을 원한다면, 그만큼 학습 데이터 수는 증가한다. 이는 계산량을 증가시킬 뿐 아니라, 인식 결과에도 악영향을 미친다.

두 번째, 다른 행동임에도, 경우에 따라 유사한 움직임 템플릿이 생성되는 경우이다. 앞으로 걸어 나오는 동작과 제자리에서 몸을 좌우로 흔드는 동작, 손을 앞으로 내미는 동작과 좌우로 손을 흔드는 동작, 그리고 고개를 앞으로 움직이는 인사하기 동작과 고개를 좌우로 갸우뚱거리는 움직임 동작 등은 서로 다른 동작임에도 불구하고 정면 카메라에서 바라볼 때 유사한 모션 템플릿이 만들어 진다. 이러한 모션 템플릿이 유사한 동작은 인식 시스템에서 거짓 수락을 발생시키고, 성능 저하의 원인이 된다. 이 같은 모션 히스토리 이미지의 문제점은 광학 축(optical axis)에 평행한 모션 정보를 담지 못한다는 데에서 발생한다. 앞서 언급한 문제가 되는 행동들은 좌우의 움직임 보다 앞뒤의 움직임이 더 중요한 경우이기 때문이다. 따라서 모션 히스토리 이미지를 이용한 제스처 인식 시스템은 제스처의 방향이 시각 축에 수직인 경우에 가장 좋은 결과를 내지만, 그 외의 경우에서는 확장성이 제한된다.

3. 볼륨 모션 템플릿

3.1 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. 또한 식(1)과 같이 3.3절에서 설명할 복원 객

체의 모멘트를 이용하여 중심점을 계산하였다.

$$m_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q z^r R(x,y,z)dxdydz \quad (1)$$

그러면, 중심점 $(\bar{x},\bar{y},\bar{z})$ 는 아래와 같이 얻어진다

$$(\bar{x},\bar{y},\bar{z}) = \left(\frac{m_{100}}{m_{000}}, \frac{m_{010}}{m_{000}}, \frac{m_{001}}{m_{000}}\right) \quad (2)$$

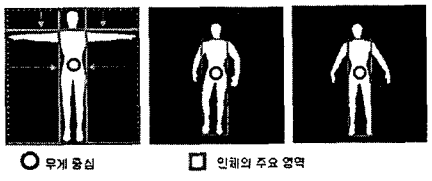

그림 1. 공간 정규화 결과

3.2 시간 정규화

제스처의 속도는 다양한 사람뿐 아니라 같은 사람의 행동 시마다 편차를 갖는다. 이러한 점은 형상기반 인식 시스템의 성능에 악영향을 끼친다. 모션 히스토리 이미지는 제스처의 최소, 최대 기간을 정해 그 사이의 모든 템플릿을 입력 영상마다 모두 생성하는 방법을 제안하지만 비효율적이다. 또한 차후 발표된 tMHI에서는 타임스탬프값을 이용하는데, 시스템간의 성능편차만 줄일 뿐 수행시마다 발생하는 제스처의 속도 편차는 수용하지 못한다[3]. 그림 2 (a)와 (b)를 보면, 같은 동작임에도 제스처의 속도 차이로 인해 다른 모션 템플릿이 생성된 것을 볼 수 있다. 이를 해결하기 위해 모션량을 정의하고 이를 이용하여 시간흐름에 따른 모션 히스토리 정보의 사라짐을 방지하여, 시간 정규화를 하였다. 연속되는 두 복원객체의 차이 D를 아래와 같이 정의하고,

$$D_t(x,y,z) = \left|R_t(x,y,z) - R_{t-1}(x,y,z)\right| \quad (3)$$

모션량 μ_t 를 다음과 같이 계산한다

$$\mu_t = \iiint D_t(x,y,z)dxdydz \quad (4)$$

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

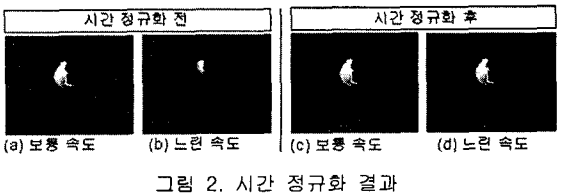

그림 2. 시간 정규화 결과

3.3 볼륨 모션 템플릿

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 같다.

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

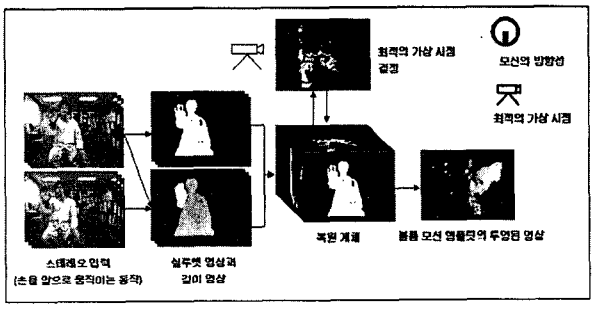

그림 3. 볼륨 모션 템플릿 생성을 위한 절차

복원된 객체는 실루엣 영상과 깊이영상을 통해 아래와 같이 정의되어 진다.

$$R_t(x,y,z) = \begin{cases} 1 & \text{if } S_t(x,y) = 1 \text{ and } \Omega Z_t(x,y) = z \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

여기서 S_t 는 실루엣 이진 영상, Z_t 는 깊이 영상 그리고 Ω 는 깊이 영상의 정규화를 위한 상수이다.

연속되는 두 R_t의 차이인 D_t를 통해 시간 t의 VMT는 다음과 같이 생성된다.

$$V_t(x,y,z) = \begin{cases} i_{\max} & \text{if } D_t(x,y,z) = 1 \\ \max(0, V_{t-1}(x,y,z) - \eta\mu_t) & \text{otherwise} \end{cases} \quad (6)$$

여기서 $i_{\max}$는 명암의 최대값 그리고 η는 히스토리 정보 감소를 위한 상수이다.

VMT는 3차원 공간에 표현되었으므로, 단순히 Y축 회전만으로 시점을 바꿀 수 있다. 최적의 가상 시점을 찾기 위해 우리는 상단 시점에서 내려다 본 투영 이미지 $\tilde{\rho}_v$를 이용한다.

$$\tilde{\rho}_v(x,y) = \mathrm{proj}_{xz} V_v(x,y,z) \quad (7)$$

상단 시점에서 움직임의 주요 방향 벡터를 구한다. 움직임을 가장 잘 표현할 수 있는 시점은 움직임의 방향과 수직인 곳이다.

크기가 3x3인 소벨 연산자를 이용하여 상단 시점의 투영영상 $\tilde{\rho}_v$의 모션 방향 $\bar{\phi}$을 구한다. 모션 방향 $\bar{\phi}$의 수직인 곳이 최적 가상 시점으로 결정되며, VMT를 회전하고 투영하여 시점에 독립적인 모션 템플릿을 생성한다.

$$\rho_v(x,y) = \mathrm{proj}_{xy}(\Psi_{\bar{\phi}} V_v(x,y,z)),$$

$$\Psi_{\bar{\phi}} = \begin{bmatrix} \cos\bar{\phi} & 0 & -\sin\bar{\phi} & 0 \\ 0 & 1 & 0 & 0 \\ \sin\bar{\phi} & 0 & \cos\bar{\phi} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

여기서 $\Psi_{\bar{\phi}}$는 VMT의 가상 시점 변환을 위한 회전 행렬이다.

다양한 시점에서의 입력영상에 무관하게 모션 방향을 기준으로 최적 가상 시점을 결정함으로써 VMT는 시점에 독립적으로 일관성 있는 템플릿을 보여준다.

4. 실험 결과 및 분석

실험을 위해 Videre Design사의 STH-MDCS2 카메라에 4.8mm 렌즈를 장착하여 사용되었다. 입력영상의 해상도는 320x240 24bit RGB 영상이었고, 초당 24장 이상의 영상 획득 환경에서 실험하였다.

제안된 VMT의 효용성을 테스트하기 위해 손앞으로 움직이기 행동에 대하여 0°~90°사이의 7가지 다른 시점에서 획득한 입력 영상에 대하여 실험하였다. 그림 4와 그림 5는 VMT와 MHI를 비교한 결과이다. 그림과 그래프를 통해 알 수 있듯이, 기존 방법에서는 입력 시점이 변함에 따라 모션정보를 담고 있는 템플릿이 다양하게 변하며, 유사도 또한 떨어지는 것을 알 수 있다. 그림 5는 시점 0°를 기준으로 하여 다른 시점 입력 결과와의 상관계수를 계산한 결과이다. MHI는 시점이 변함에 따라 민감하게 반응하였고, 이는 같은 행동이라도 카메라 시점이 변함에 따라 인식결과에 영향을 끼치며, 시점별 템플릿을 학습시켜야 하는 점을 보여준다. 반면, 제안된 방법에서는 시점이 다양하게 변하여도 유사성 있는 모션 템플릿이 생성됨을 알 수 있었다. 이는 시점에 독립적인 제스처 인식을 할 수 있음을 의미한다.

본 논문은 시점에 독립적인 제스처 인식을 위해 스테레오 정보를 활용한 볼륨 모션 템플릿을 제안하였다. 다양한 시점 입력에서의 제스처 인식을 위해 기존 연구에서는 각 시점별로 학습 데이터가 필요하였고, 그에 따른 계산 량과 인식 성능이 저하되는 단점이 있었지만, 제안된 방법은 깊이 정보를 활용하여 3차원 공간상의 모션 정보를 복원하였을 뿐 아니라, 최적의 시점을 찾아 모션이 가장 잘 표현되는 템플릿을 생성할 수 있게 하였다.

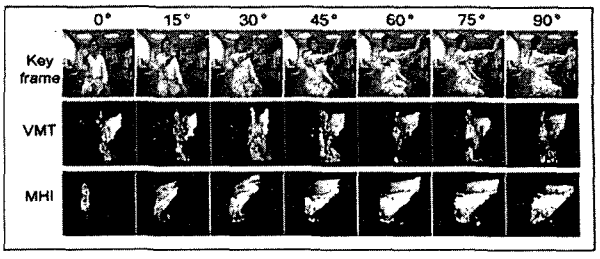

그림 4. 다양한 시점 입력에 대한 VMT와 MHI 결과 비교

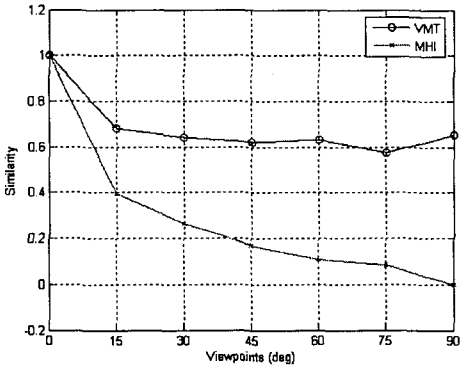

그림 5. 시점에 따른 템플릿 유사도

또한 제스처가 발생할 때마다 생기는 속도 편차 문제를 위해 모션량을 계산하고 이를 통해 히스토리 정보를 유동적으로 갱신하는 새로운 시간 정규화 방법을 제안하였다.

향후 연구로는 다양하고 복잡한 동작들로 볼륨 모션 템플릿을 검증하고, 데이터베이스를 통해 인식 실험하여 보다 시점에 강인한 인식 시스템을 구현하려고 한다.

참고 문헌

[1] 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, March 2001.

[2] 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, June 2003.

[3] G. R. Bradski and J. W. Davis, "Motion Segmentation and Pose Recognition with Motion History Gradients," Machine Vision and Applications, Vol. 13, pp. 174-184, 2002.

[4] G. Ye, J. Corso and G. Hager, "Gesture Recognition Using 3D Appearance and Motion Features," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2004.

[5] R. Bodor, B. Jackson, O. Masoud and N. Papanikolopoulos, "Image-Based Reconstruction for View-Independent Human Motion Recognition," Proc. of the IEEE conference on Intelligent Robots and System, Las Vegas, pp. 1548-1553, October 2003.