깊이 영상에서 외형 기반 파티클 필터를 이용한 3 차원 손 포즈 추정

유수곤[†, O], 신봉기[‡], 이성환[†]

[†]고려대학교 뇌공학과 [‡]부경대학교 IT 융합응용공학과

sgriew@image.korea.ac.kr bkshin@phnu.ac.kr swlee@image.korea.ac.kr

3D Hand Pose Estimation using an Appearance-based Particle Filter from Depth data

Soo-Gon Riew[†,O], Bong-Kee Shin[‡], Sung-Whan Lee[†]

[†] Dept. of Brain and Cognitive Engineering, Korea University

[‡] Dept. of IT Convergence and Application Engineering, Pukyong National University

요 약

본 연구에서는 Kinect에서 추출된 깊이 영상을 이용하여 고차원 공간에서 3차원 손 포즈 추정을 위해 샘플링 방법을 개선한 외형 기반 파티클 필터를 제안한다. 제안된 방법은 파티클 필터의 샘플로 사용될 전역적 손 포즈 추정을 위해 주성분 분석과 자기 조직화 지도를 이용한 클러스터링 방법으로 3차원 손 모델의 대략적인 손 관절의 포즈를 추정한다. 전역적 손 포즈는 이전 영상에서 추정된 샘플의 동적 모델과 결합하여 파티클 필터의 샘플을 생성하고 각 포즈의 파라미터 추정을 위해 분포한다. 입력 영상의 3차원적인 특징을 이용하여 관측 모델로 입력 영상과 3차원 모델에서 추출한 깊이 정보를 거리 매칭함으로써 큰 깊이값 차이가 생기는 포즈를 잘못 추정된 포즈로 배제시킨다. 실험 결과, 손목의 회전과 손가락의 가려짐 현상이 발생한 경우에도 강인한 추적 결과를 제공함을 확인할 수 있다.

1. 서 론

컴퓨터 비젼 분야에서 3차원 손 포즈 추정 연구는 제스처 인식을 이용한 게임 인터페이스, 지시형 제스처, 수화 인식 등의 여러 분야에서 핵심적인 요소로 사용될 수 있다. A. Erol 등은 손 포즈 추정 방법을 크게 '외형 기반 방법'과 '모델 기반 방법'으로 나눈다[1]. '외형 기반 방법'은 영상의 특징을 추출하여 학습 데이터를 생성 후 인식 알고리즘으로 학습하고 입력 데이터의 특징과 비교하여 포즈를 추정하는 방법이다. 이 방법은 입력 데이터에서 손을 추정하는데 초기화 단계가 필요 없고, 적은 계산량을 요구하지만 학습 데이터에 입력 데이터와 일치하는 포즈가 없다면 잘못된 포즈를 추정하는 오류가 발생하게 된다. 반면, '모델 기반 방법'은 원통, 구, 타원 등을 이용하여 모델을 정의한 후, 각 파라미터를 확률적으로 추정하여 포즈를 표현한다. 이 방법은 정교한 포즈 추정이 가능하지만 고차원 공간에서 포즈를 추정해야 하기 때문에 계산량이 많은 단점이 있다.

위의 두 방법을 결합하여 B. Stenger 등은 에지, 컬러 우도를 관측 모델로 하여 Tree-based filtering을 통해 외형 기반의 추정을 하고, 추정된 파라미터를 이용하여 3차원 모델을 재구성하였다[2]. 하지만 외형 기반에 의존적이어서 추정 가능한 포즈가 제한적이다. Y. Wu 등은 주성분 분석으로 차원을 축소한 기저 이미지를 이용하여 포즈의 외형을 추정을 한 뒤, Monte Carlo 알고리즘을 이용하여 cardboard 모델과 입력 영상과 비교하여 추정하였다[3]. 하지만 손을 정면에서 바라보았을 경우에만 추정이 가능하고, 이전 영상에서 추정된 결과와 현재 영상의 정보만을 사용하기 때문에 국소 최저치에 빠질 위험이 있다는 단점이 있다. 최근에 손 포즈 측정을 위해 C. Weng 등은 파티클 필터를 개선한 Separable State Based Particle Filtering(SSBPF)를 제안하였다[4]. SSBPF는 비 정상적인 손가락의 운동학(kinematics)을 추정하기 위해 각 손 관절을 독립적인 상태로 나누고 각 관절에 적응적으로 파티클을 분포함으로써 계산량을 줄였다. 하지만 손이 정면을 바라봐야 하는 제약이 있다.

본 연구에서는 손의 포즈 제한과 국소 최저치에 빠지는 문제를 해결하고, 회전 및 손가락 간의 가려짐에도 강인한 포즈 추정 방법을 제안한다. 제안된 방법은 자기 조직화 지도(Self-Organizing Map; SOM)[5]에서 추정된 전역적 손 포즈와 이전 영상에서 관측된 샘플의 동적 모델을 결합하여 외형 기반 파티클 필터의 샘플링 단계를 향상시킨다. 향상된 샘플링 단계는 손의 추정이 과도하게 잘못 추정되거나 국소 최저치에 빠지는 것을 방지한다. 결합된 샘플은 3차원 모델과 입력 영상의 깊이 정보를 이용한 거리 매칭 방법과 파티클 필터를 이

용하여 지역적 손 포즈를 추정함으로써 가려짐에도 양호한 결과를 보였다.

본 논문의 2장에서는 SOM 클러스터 학습을 위한 특징 벡터 생성, 학습, 외형 기반 파티클 필터에 대해 설명하고 3장에서는 이를 이용한 포즈 추정 실험 결과 및 분석을 보이고, 4장에서는 결론 및 향후 연구에 대해 서술한다.

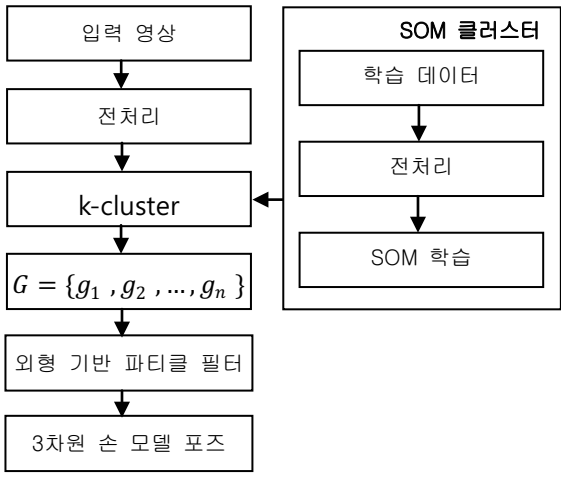

그림 1 시스템 기능 및 구조도

2. 제안하는 손 포즈 추정 방법

본 연구에 대한 전반적인 시스템의 구성은 그림 1에 나타나 있다. 시스템은 크게 두 부분으로 나뉜다. 첫 번째로 입력 영상이 전처리를 거쳐 특징 벡터로 변환되고 미리 학습된 SOM 클러스터의 입력값으로 사용된다. SOM 클러스터는 전역적인 포즈 벡터를 출력하고 3차원 모델의 초기정보로 사용된다. 두 번째로 전역적인 포즈 벡터를 기반으로 외형 기반 파티클 필터를 이용하여 정교한 포즈 추정을 한다. 이때, 분할된 손 영상의 깊이 정보와 에지, 3차원 손 모델의 깊이 정보와 에지의 거리를 관측값으로 거리 매칭을 하여 추정 오차를 측정한다.

2.1 특징벡터 생성과 SOM 클러스터 학습

손 영역 분할 및 학습 데이터 생성

손 영역은 손의 위치에서 주어진 깊이에 임계값을 주어 뒷배경을 제거하고, 수식 (1)을 이용하여 H_r 크기만큼 가로 S_x, 세로 S_y를 분할한다.

$$\begin{aligned} h_x - H_r < S_x < h_x + H_r \\ h_y - H_r < S_y < h_y + H_r \end{aligned} \quad (1)$$

h_x , h_y는 손의 중심 좌표를 나타낸다. H_r은 손의 중심 좌표의 깊이 정보에 따라 유동적으로 변하며 실험적으로 주어졌다. 분할된 손 이미지는 같은 크기 영상으로 정규화하고 깊이 형상을 강조하기 위해 명암비를 조절한다.

SOM의 입력 벡터를 생성하기 위해 주성분 분석법으로 차원을 축소한다. 주성분 분석은 영상을 축소할 때 패턴 정보를 최대한 확보하고 의미 있는 정보 패턴을 유지하여 차원을 축소한다.

SOM 클러스터의 결과값인 전역적 포즈는 깊이 영상과 함께 획득된다. 본 연구에서는 각 학습 데이터의 손 관절의 각도에 대한 정보를 획득하기 위해 Vicon의 모션 캡처 시스템을 이용하였다.

Kohonen의 SOM

SOM은 자기 조직화의 특성을 이용한 무 감독형 클러스터이다. SOM은 입력층과 출력층으로 나뉘는데 입력되는 학습 데이터는 학습 단계에서 출력층의 미리 정해진 크기의 노드와 완전 연결이 되며, 각 노드의 가중치를 학습하고 최종적으로 2차원의 출력 노드 집합으로 양자화시킨다. SOM은 가중치를 주는 과정에서 전방 패스(feedfoward)를 사용하는 구조이며 학습된 결과는 위상적 특징을 가지고 있다. 즉, 인식 수행이 상당히 빠르고, 한 노드와 주위 노드에 비슷한 특징을 가진 데이터가 자리하게 된다. 이와 같은 특징을 이용하여 전역적 포즈를 추정하는데 SOM을 이용하였다.

2차원으로 양자화된 출력층의 각 노드는 전역적인 포즈 벡터 $G = \{g_1, g_2, \ldots, g_n\}$로 이루어져 있다. n은 관절의 수를 나타낸다. SOM이 입력 영상의 클러스터로 사용됨으로써 손 관절 포즈를 전역적 포즈 추정을 할 수 있고, 이를 통해 3차원 손 모델 초기화의 문제를 해결하고 파티클 필터의 샘플링을 개선할 수 있다.

2.2 외형 기반 파티클 필터를 이용한 3차원 손 포즈 추론

3차원 손 모델링

손 모델은 I. Oikonimidis 등 [6]에서 소개하고 있는 모델을 참고하였다 (그림 2(a)).

손 모델은 손바닥과 손가락의 움직임으로 구성되어있다. 손바닥의 움직임 M_{Palm}은 3차원 전이행렬 T와 3차원 회전행렬 R로 이루어지고 자유도는 6이다. 손가락 M_{Finger}의 관절에 대한 자유도는 총 19이고, 다음과 같이 표현한다.

$$\begin{aligned} &M_{finger} = \{\theta_j^i; i, j\}, \\ &i = \{Thumb, Index, Middle, Ring, Little\},\ j = \{near, middle, far\} \end{aligned} \quad (2)$$

여기서 i는 손가락을 나타내고, j는 관절을 표현한다. θ는 각 관절의 각도를 나타낸다. 각 손가락은 상호 간의 의존관계를 가지는데 그림 2(b)과 같은 방법으로 자유도를 줄일 수 있다.

외형 기반 파티클 필터

손 포즈의 추정을 해결하기 위해 외형 기반 파티클 필터 알고리즘을 제안한다. 외형 기반 파티클 필터는 SOM에서 추정된 G와 이전 영상에서 추정된 X_{t-1}의 동

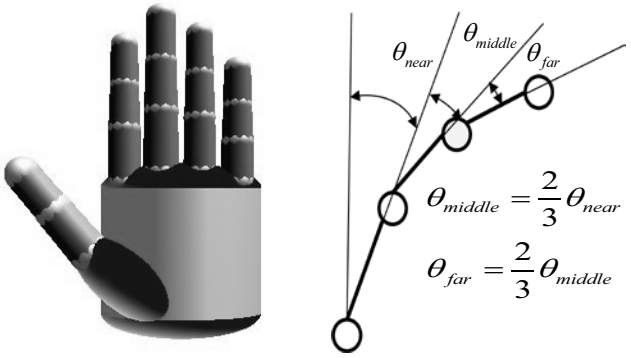

그림 2 (a) 3차원 손 모델 (b) 손가락 관절 각도 표현

적 모델을 기반으로 지역적 손 움직임 $X = \{x_1, x_2, \ldots, x_n\}$을 추정한다. n은 관절의 수를 나타낸다. 이를 베이지안 룰 에 따라 수식화 하면 식(3)와 같다.

$$p(X_t \mid Y_t, G_t) = \frac{p(Y_t \mid X_t, G_t)\, p(X_t \mid Y_{1:t-1}, G_t)}{p(Y_t, G_t)} \quad (3)$$

$p(Y_t|X_t, G_t)$ 는 우도이고 t는 시간을 나타낸다. $p(X_t|Y_{1:t-1}, G_t)$는 사전 확률이며, 식(4)와 같이 나타낼 수 있다.

$$p(X_t \mid Y_{1:t-1}, G_t) = \int_{t-1} p(X_t \mid X_{t-1})\, p(X_{t-1} \mid Y_{t-1}, G_{t-1})\, dX_{t-1} \quad (4)$$

사전 확률은 파티클의 분포 방향을 결정하며 동적 모델을 통해 구한다.

$$p(X_t \mid X_{t-1}) = \exp(-\left\| X_{t-1} - X_{t-2} \right\|) \quad (5)$$

동적 모델은 이전에 추정된 포즈에 의해서 Euclidean 거리를 사용하여 구한다.

관측 모델 및 매칭 방법

손 포즈를 추정을 위한 특징으로는 에지, 실루엣, 3차원 정보 등이 있는데, 손 포즈 추정 성능은 이러한 특징을 이용하여 3차원 모델과 입력 영상의 확률 추정에 사용되는 매칭 방법에 따라 성능이 좌우된다. 제안된 방법에서는 입력 영상의 3차원적인 특징을 이용하여 관측 모델 중 하나로 입력 영상과 3차원 모델에서 추출한 깊이 정보를 사용한다. 두 깊이 정보를 거리 매칭하면 입력 영상과 다른 포즈는 뒷배경과 겹치게 되는 부분에서 큰 차이를 가지기 때문에 큰 깊이값 차이가 생기는 장점이 있다. 깊이에 대한 거리 매칭은 Y_d로 표현한다. 또한 에지 기반의 거리 측정 방법 Y_e도 사용하여 다음과 같이 $Y = (Y_d, Y_e)$ 쌍으로 하고, 두 특징을 결합 확률로써 표현한다.

$$\begin{aligned} p(Y \mid X, G) &= p(Y_d, Y_e \mid X, G) \\ &\triangleq \alpha \cdot p(Y_d \mid X, G) + (1-\alpha) \cdot p(Y_e \mid X, G) \end{aligned} \quad (6)$$

α 는 3차원 정보를 이용한 거리 측정과 에지 기반의 거리 측정의 의존도를 결정하는 상수이다. 3차원 정보를 이용한 거리에 대한 우도는 평균이 0인 정규화 분포를 따르며 수식(7)과 같이 나타낼 수 있다.

$$p(Y_d \mid X, G) \sim N(D_d; 0, \sigma_1) \quad (7)$$

여기서 σ_1는 분산이며, $D_d = |O_d - M_d|$인 유클리디안 거리로 계산한다. O_d는 분할된 손의 깊이 정보, M_d는 3차원 모델의 깊이 정보이다. 에지 기반의 거리에 대한 우도도 3차원 정보를 이용한 거리에 대한 우도와 같이 평균이 0인 정규화 분포를 따르며, 수식(8)와 같이 나타낼 수 있다.

$$p(Y_e \mid X, G) \sim N(D_e; 0, \sigma_2) \quad (8)$$

여기서 σ_2는 분산이며 D_e는 Chamfer 거리를 사용한다.

3. 실험 결과 및 분석

3.1 실험 환경

Kinect 센서에서 OpenNI의 PrimeSense Sensor Module을 이용하여 640X480 크기, 11비트 깊이 영상을 획득했다[7]. 학습 데이터는 4명의 사람에게서 영상을 획득하였고, 손 동작은 손 오므리기, 펴기, 각 손가락 굽히기, 회전하기를 반복하였다. 데이터는 두 가지를 동시에 획득하였는데 첫 번째는 깊이 영상에서 분할된 손 영역을 같은 크기 영상으로 정규화하고 깊이 형상을 강조하기 위해 명암비를 조절한 데이터이고, 두 번째는 Vicon 사의 모션 캡처 시스템을 이용하여 손 관절에 마커를 부착한 후 각 관절의 위치를 획득하였다

3.2 실험 결과 및 분석

그림 3의 (a) 영상은 분할된 손 영역, (b) 영상은 전역적 손 추정, 그리고 (c)는 최종 추정된 3차원 모델을 보여준다. 그림 3의 1~4열 영상은 정면을 바라보고, 추정한 결과이고 5열은 손목이 회전한 결과이다. 그림 3(b)영상에서는 SOM 클러스터에서 추정된 결과로 대략적인 손 추정이 된 것을 보여준다. 그림3 (c)영상에서는 외형 기반 파티클 필터를 통해 추정된 결과로 (b)영상보다 더 정교하게 추정된 것을 보여준다. 4열 영상은 손을 오므린 포즈로 추정된 모델이 완전히 오므리지는 않았지만 비슷한 포즈로 추정되었다. 5열 영상은 손목 회전과 손가락의 움직임의 추정 결과를 볼 수 있다. 손목이 회전하면서 생기는 손가락 간의 가려짐이 발생하였음에도 불구하고 각 손가락의 위치가 추정됨을 확인할 수 있다.

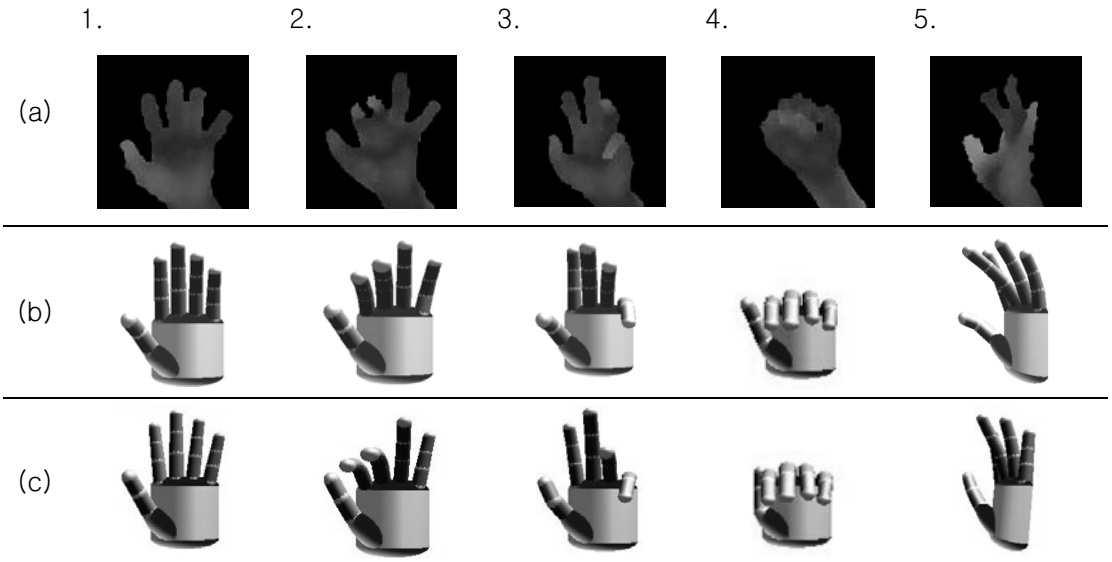

그림 3 입력 영상과 추정 결과: (a) 분할된 손 영역, (b) 전역적 추정 결과, (c) 최종 추정 결과

4. 결론 및 향후 연구

본 연구에서는 깊이 영상에서 외형 기반 방법으로 전역적 포즈를 구하고 샘플링 방법을 개선한 파티클 필터를 이용하여 3차원 손 포즈를 추정하는 방법을 제안하였다. 깊이 정보를 가진 분할된 손 이미지의 차원을 축소 한 후 SOM 클러스터를 학습하고, 입력 영상이 SOM을 통해 얻어진 레이블에서 손가락 관절 정보를 획득함으로써 전역적인 손 포즈를 추정하여 첫 프레임에서의 초기화 문제를 해결하였다. 전역적 포즈를 이전 영상에서 추정된 샘플의 동적 모델과 결합하여 손의 추정이 과도하게 잘못 추정되거나 국소 최저치에 빠지는 것을 방지하였다. 또한 깊이 정보를 관측값으로 사용함으로써 포즈가 다른 모델에 대해서 입력 영상과 큰 차이를 주어 잘못 추정된 포즈로 배제할 수 있었다. 추정 결과 손바닥이 카메라를 바라본 상태에서 손가락 간의 움직임 추정이 잘 되었고, 손이 회전하여 손가락 간의 가려짐이 발생한 경우에도 상대적으로 양호한 결과를 얻을 수 있었다. 향후 연구는 Kinect에서 함께 제공되는 컬러 영상에서 손을 추출하고, 하나의 관측 모델로 사용함으로써 포즈 추정 성능을 개선시키고자 한다. 또한 GPU 기반의 병렬처리를 적용하여 추정하는데 걸리는 시간을 단축할 수 있을 것이다.

감사의 말씀

본 연구는 한국연구재단을 통해 교육과학기술부의 세계수준의 연구중심대학육성사업(WCU)으로부터 지원받았음 (R31-10008).

참고문헌

[1] A. Erol, G. Bebis, M, Nicolescu, R. Boyle, and X. Twombly, "Vision-based Hand Pose Estimation: A Review," Computer Vision and Image Understanding, Vol. 108, No. 1, pp. 52-73, 2007.

[2] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla, "Model-Based Hand Tracking Using a Hierarchical Bayesian Filter," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 28, No. 9, pp. 1372-1384, 2006.

[3] Y. Wu, J. Lin, and T. Huang, "Capturing Natural Hand Articulation," IEEE Proc. International Conference on Computer Vision, Vol.2, pp.426-432, 2001.

[4] C. Weng, C. Tseng, M. Ho, and C. Huang, "A Vision-based hand motion parameter capturing for HCI," Proc. IEEE International Conference on Audio, Language and Image Processing, pp. 1219-1224, 2008.

[5] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, 2001.

[6] I. Oikonomidis, N. Kyriazis, and A. Arguros, "Markerless and Efficient 26-DOF Hand Pose Recovery," Asia Conference on Computer Vision, pp.1-13, 2010.

[7] OpenNI. PrimeSense Sensor Module, 2011. URL http://github.com/PrimeSense/Sensor