ETRI

2014
Electronics and
Telecommunications
Trends

차세대 콘텐츠기술 특집

거리영상 기반 동작인식 기술동향

Technology Trends of Range Image based Gesture Recognition

장주용 (J.Y. Chang) 인터랙티브입체영상연구실 선임연구원
류문욱 (M.W. Ryu) 인터랙티브입체영상연구실 연구원
박순찬 (S.C Park) 인터랙티브입체영상연구실 연구원

* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 콘텐츠산업기술 지원 사업으로 수행되었음 (APP0120130417001 인터랙티브 콘텐츠와 상호작용을 위한 고정밀 모바일 파노라믹 360도 다수 사용자 동작인식 기술개발).

동작인식(gesture recognition) 기술은 입력 영상으로부터 영상에 포함된 사람들의 동작을 인식하는 기술로써 영상감시(visual surveillance), 사람-컴퓨터 상호작용(human-computer interaction), 지능로봇(intelligence robot) 등 다양한 적용분야를 가진다. 특히 최근에는 저비용의 거리 센서(range sensor) 및 효율적인 3차원 자세 추정(3D pose estimation)기술의 등장으로 동작인식은 기존의 어려움들을 극복하고 다양한 산업분야에 적용이 가능할 정도로 발전을 거듭하고 있다. 본고에서는 그러한 거리영상(range image) 기반의 동작인식 기술에 대한 최신 연구동향을 살펴본다.

Ⅰ. 서론

2차원 영상으로부터 사람의 동작을 인식하는 연구는 컴퓨터비전(computer vision)의 초창기부터 수행되어 온 매우 중요한 연구분야 중의 하나로써 영상감시(visual surveillance), 사람-컴퓨터 상호작용(human-computer interaction), 지능로봇(intelligent robot) 등 다양한 적용분야를 가진다. 동작인식에서 인식의 대상인 사람의 동작은 다양한 의미를 지닐 수 있는데, 신체부위들이 어떻게 배치(a configuration of the human body)되어 있는가를 표현하는 자세 혹은 특정한 의미를 가지는 신체의 움직임을 나타내는 동작(gesture) 등을 들 수 있다. 자세의 경우 비교적 변형되지 않는(rigid) 신체부위들의 위치와 그 부위들 사이의 연결정보를 기반으로 표현하는 스켈레톤(skeleton)이 대표적이다. 이는 어떠한 의미도 배제한 채 사람의 형태정보만을 추상적/효율적으로 전달하는 데에 그 목적이 있다. 동작의 경우 일반적으로 신체부위들의 동적인 움직임을 가정하나 정적인 자세(posture) 또한 넓게 보아 짧은 시간간격을 가지는 동작의 일부로 간주할 수 있다. 이러한 동작은 행위자가 의식적 혹은 무의식적으로 전달하고자 하는 특정한 의미를 담고 있는 것이 일반적이다. 이 중 본고에서는 두 번째 의미의 동작인식을 다룬다.

사람의 경우 두 눈을 통해 대상자의 동작을 큰 어려움 없이 인식(recognition)하고 지각(perception)할 수 있는데 반해, 컴퓨터로 하여금 입력 컬러영상으로부터 그러한 동작인식 작업을 자동으로 수행하도록 만드는 것은 결코 쉽지 않다. 2차원 컬러영상으로부터의 동작인식을 어려운 문제로 만드는 요인들로는 다음과 같이 여러 가지를 들 수 있다. 첫째, 사람이 취할 수 있는 자세 및 동작은 수학적으로 고차원이며, 따라서 동작인식과 관련된 탐색 및 최적화과정을 매우 복잡하게 만든다. 두 번째로 동일한 동작이라고 하더라도 그것을 수행하는 사람에 따라서, 심지어는 동일한 사람의 경우에도 수행되는 동작이 매번 유사하지 않으며, 때로는 매우 큰 변동을 가질 수 있다. 세 번째, 사람의 동작은 기본적으로 3차원 공간에 속하므로 2차원 영상으로의 투사(projection)과정에서 정보의 손실이 발생하게 된다. 또한 배경과 사람과의 분리 문제 그리고, 사람의 컬러 혹은 텍스처에 해당하는 외양(appearance)정보의 극심한 변동 가능성 등을 들 수 있다.

한편 최근에는 기존의 컬러영상과는 다른 거리 영상(range image)에 기반한 동작인식 기술들이 실용화되고 있다. 대표적으로 Microsoft의 Kinect[1)]를 포함하여 최근 등장하고 있는 저가의 거리 센서들은 대상 장면에 대한 3차원 정보를 실시간(30 frame per second)으로 제공한다. 이는 배경과 사람과의 분리를 용이하게 하며, 사람의 외양 정보에 의존하지 않는 3차원 거리 영상의 활용을 가능하게 하여 기존 동작인식의 어려움들을 상당 부분 극복할 수 있도록 해 준다. 또한 최근 활발히 연구되고 있는 거리 영상 기반의 3차원 자세 추정(3D pose estimation) 방법들 또한 3차원 거리 센서와 함께 대상 사람의 체형 변화에 무관하게 사람의 3차원 자세를 스켈레톤의 형태로 효과적으로 추출할 수 있도록 해 주는데[1], 이러한 스켈레톤과 같은 중간단계 특징(mid-level feature) 정보의 사용은 거리영상 기반 동작인식 기법들의 실용화를 가능하게 만든 주요 요인이다.

본고에서는 최근 활발히 연구되고 있는 거리영상 기반의 동작인식 기술 연구동향을 살펴본다. Ⅱ장에서는 거리영상의 획득을 가능하게 해 주는 거리 센서동향을, Ⅲ장에서는 입력 거리영상으로부터 대상 사람의 전신 동작을 인식하는 기술동향을, Ⅳ장에서는 사람의 신체부위 중 메시지 전달에 있어 가장 큰 비중을 차지하는 손 동작인식동향을 기술한다. 마지막으로 Ⅴ장에서는 결론 및 향후 발전방향을 기술한다.

---

1) http://www.microsoft.com/en-us/kinectforwindows/

II. 거리 센서기술

거리 센서는 원거리에서 3차원 데이터를 획득할 수 있는 특징을 기반으로 제품 검사, 물체/환경 인식 및 모델링, 역공학, 동작인식 분야에서 널리 사용되고 있다[2]. 거리 데이터의 획득을 위해서 극초단파(microwave), 광파(light wave), 초음파(ultrasonic wave)를 사용할 수 있으며, 광파를 이용한 방식은 크게 삼각 측량(triangulation), 간섭 측정(interferometry), 시간 지연 측정(time-of-flight)으로 나눌 수 있다. 삼각 측량 방법은 두 개 이상의 좌표계가 가지는 상호 간의 거리와 측정하고자 하는 물체와의 각도들을 이용하여 거리 데이터를 획득하는 방법이고, 시간 지연 측정방법은 물체에 투사된 광파가 되돌아 오는 시간을 측정하여 거리 데이터를 획득하는 방법이다. 간섭 측정방법은 시간 지연 측정법과 원리는 유사하나 물체에서 반사되는 빛의 시간 차이를 광학적 간섭계를 이용하여, 기준파와의 위상 차이를 측정하여 거리 데이터를 획득한다[3].

본 장에서는 동작인식 기술에 주로 사용되는 삼각 측량, 시간 지연 측정 기반의 거리 센서기술 동향에 대해 기술한다.

1. 삼각 측량 기반 거리 센서

삼각 측량 기반 거리영상 센서는 인간의 두 눈과 같이 특정한 베이스라인(baseline)을 가지는 두 좌표계에서 영상을 획득하고, 두 영상에서의 대응점을 찾아 거리를 계산한다. 대응점을 찾아야 거리영상을 복원할 수 있기 때문에 텍스처가 없는 환경이나 조명변화가 심한 환경에서는 사용하기 힘든 단점이 있다. 이를 보완 하기 위해 하나의 카메라 대신 광파를 이용하기도 하며, 광파의 사용 유무에 따라 수동(passive)방식과 능동(active)방식으로 나눌 수 있다.

가. 스테레오

스테레오 비전(Stereo vision)의 경우 빠른 속도로 대응점을 찾거나 텍스처가 부족한 환경에서 매칭을 시도하는 방법 등 효율적으로 스테레오 매칭을 시도하는 연구가 수행되어 왔으며[4], 여러 시점의 영상을 기반으로 거리영상을 복원하는 방법[5], 텍스처가 부족한 환경을 보완하기 위해 임의 패턴을 추가하여 능동 스테레오(active stereo)를 사용하는 방법 등 여러 방면에서 연구가 진행되고 있다. 상용 스테레오 카메라 제품으로는 Point Grey Research의 Bumblebee[2] 제품이 있다.

나. 구조광

구조광은 대표적인 능동방식의 삼각 측량 방법으로, 스테레오 카메라에서 하나의 카메라를 패턴을 투사할 수 있는 프로젝션 시스템으로 대체한 형태로 구성된다. 패턴을 투사하기 때문에 텍스처가 없는 환경에서도 거리를 계산 할 수 있으며, 미리 정의된 패턴으로 하나의 카메라 영상에서 대응점을 찾을 수 있다. 때문에 대부분의 연구는 목적에 알맞은 패턴설계를 통한 거리영상 획득속도, 정확도/정밀도 향상, 혹은 환경 강인성에 대한 연구들에 집중되어 왔다[6]. 구조광 방법은 패턴 코딩방법에 따라 크게 직접 코드(direct code), 공간 코드(spatial code), 시간 코드(temporal code)로 나눌 수 있다. 시간 코드는 가장 높은 정확도/정밀도를 보여 주지만 여러 장의 패턴을 사용하기 때문에 속도가 느린 측면이 있고, 공간 코드는 한 장의 코드를 사용하여 속도는 빠르나 정확도/정밀도가 떨어지는 단점이 있으며, 직접 코드는 환경에 영향을 많이 받는 단점이 있다. 상용 제품으로는 Microsoft의 Kinect가 있다.

2. 시간 지연 측정 기반 거리 센서

시간 지연 측정(time-of-flight)기법은 특정 광파를

[2] http://ww2.ptgrey.com/

쏘고 되돌아 오는 시간을 측정하여 거리정보를 복원하는 방법으로 디지털 카운트를 가지고 레이저 펄스가 되돌아 오는 시간을 측정하거나 수신된 펄스의 위상 차이를 이용하여 시간을 측정하는 방법을 사용한다. 이때 서로 다른 위상 차이를 만들기 위해서 하나의 픽셀에서 노출구간을 조절하거나, 인접한 픽셀에서 서로 다른 구간 시간을 지정하는 방법 등을 사용한다[7]. 기하학적인 위치 변이를 사용하는 삼각 측량과는 달리 광파를 직접 활용하기에 이에 대한 오차를 보상하는 방법들이 연구되고 있다[8]. 상용제품으로는 Softkinetic의 DS series[3], Mesa Imaging의 SR series[4] 등이 있다.

Ⅲ. 전신 동작인식 기술

거리영상으로부터의 동작인식 기술은 두 가지 방식으로 크게 나눌 수 있는데, 즉 동작의 시간적인 변화를 동적으로 모델링하고 이 모델을 동작인식에 활용하는 순차적 접근법(sequential approach)과 시공간 안에서 이루어지는 동작을 전체적으로 파악하고 그 안에서 특징정보를 추출하여 동작인식에 활용하는 시공간적 접근법(space-time approach)이 그것이다. 본 장에서는 거리영상 기반의 전신 동작인식 기술들을 이러한 기준에 따라 분류하고 각각에 대해 간략히 소개한다.

먼저 거리영상 기반의 동작인식 방법들의 평가를 위해 사용되어 온 대표적인 데이터셋(dataset)을 간략히 언급하면 다음과 같다. 먼저 참고문헌 [9]에서는 10명의 대상자로부터 취득된 20개의 동작유형에 대한 총 567개의 거리 동영상으로 이루어진 MSR Action 3D dataset이 제안되었다. 해상도는 320x240이며 Kinect와 유사한 거리 센서가 사용되었고, 20개의 관절(joint)을 가진 스켈레톤 및 전경 사람 영역에 해당하는 영역(segmentation) 정보가 함께 제공된다. 한편 참고문헌 [10]에서는 10명의 대상자로부터 획득된 16개의 동작유형에 대한 총 320개의 거리 동영상으로 이루어진 MSR Daily Activity 3D dataset이 제안되었는데, 이는 추출된 스켈레톤의 정확도가 떨어지고, 사물과의 상호작용을 포함하며, 전경 영역을 제공하지 않는다는 측면에서 좀 더 난이도가 높은 데이터셋이라고 할 수 있다. 마지막으로 참고문헌 [11]에서는 컬러, 깊이(depth), 스켈레톤, 그리고 오디오(audio)의 다양한 modality를 가지는 동작인식 데이터셋인 ChaLearn Multi-modal Gesture dataset이 공개되었다. 이는 27명의 대상자로부터 수행된 20개의 동작으로 구성된 총 13,858개의 샘플들로 구성되어 있다. 특징적으로 이 데이터셋은 이탈리아에서 사용되는 문화적/인류학적 제스처들로 구성되어 있다.

1. 순차적 접근법

먼저 참고문헌 [12]에서는 특징정보를 추출하기 위해 영역화된 깊이 영상에 R transformation을 적용하였다. 이를 통해 얻어진 1차원 특징 프로파일(profile)은 변위와 스케일에 불변하는(translation and scale invariant) 장점을 가진다. 여기에 추가적으로 PCA(Principle Component Analysis) 및 LDA(Linear Discriminant Analysis)를 적용하여 최종 특징 벡터를 생성한다. 동작의 분류(classification)에는 discrete HMM(Hidden Markov Model)이 사용되었다. 참고문헌 [13]은 깊이 영상으로부터 얻어진 중간단계 특징이라고 할 수 있는 스켈레톤 정보를 활용하여 Histogram of 3D Joints (HOJ3D) 특징을 제안하였다. 이는 스켈레톤의 루트(root) 관절을 기준으로 다른 관절들의 공간적 분포를 히스토그램(histogram) 형식으로 계산한 것이다. 하나의 HOJ3D 벡터는 하나의 자세(posture)를 나타내며, 추가적인 LDA 및 K-means clustering을 통해 벡터 양자화(vector quantization)가 되어 최종적으로 discrete HMM을 통해 동작이 분류된

[3] http://www.softkinetic.com/en-us/softkinetic.aspx
[4] http://www.mesa-imaging.ch/index.php

다. 참고문헌 [14]에서는 깊이 영상과 스켈레톤 정보를 동시에 사용하는 하이브리드(hybrid) 특징 벡터를 제안하였는데, 이는 기존의 스켈레톤 관절의 방향(orientation), 위치(position), 움직임(motion) 정보와 함께 깊이 영상으로부터 얻어진 HOG(Histogram of Oriented Gradients) 특징정보의 결합으로 이루어진다. 또한 이 논문에서는 사람의 동작이 계층적(hierarchical) 구조를 가지고 있다는 점에 주목하여 동작을 two-layer MEMM(Maximum Entropy Markov Model)로 모델링하는 방법을 제안하였다. 참고문헌 [10]에서도 하이브리드 특징을 사용하였는데, 관절 쌍(pair)의 상대적인 위치 차이를 이용하여 스켈레톤 특징을 구성하였고, 추가적으로 특정 관절 주위의 깊이 외양정보라고 볼 수 있는 LOP(Local Occupancy Patterns)를 정의하여 특징 벡터로 사용하였다. 또한 단순히 모든 스켈레톤 관절을 모두 사용하는 것이 아니라, 동작 분류에 도움이 되는 관절의 조합을 actionlet이라 정의하고 이를 발견하고, 결합하는 방법을 제안하였다. 그리고 Fourier Temporal Pyramid 특징 벡터를 구성함으로써 별도의 학습과정 없이 동작의 동적인 모델링을 수행하였다. 참고문헌 [15]에서는 오디오 정보와 스켈레톤 정보를 함께 사용하는 multi-modal 방법을 제안하였다. 이 논문에서 다룬 데이터셋은 동작들이 한 동영상 안에서 연속적으로 이루어지므로 동작인식 방법은 동작의 유형을 분류해야 할 뿐만 아니라 발생한 위치 또한 추정해야 하는데, 즉 단순한 동작 분류가 아닌 동작 검출(detection)을 다룬다. 이를 위해 에너지 기반의 thresholding 방법을 통해 동작이 발생한 구간(interval)의 후보를 생성한다. 그리고 각각의 후보구간에 대해 오디오 및 스켈레톤 기반의 두 분류기(classifier)를 동시에 적용한다. 오디오의 경우 MFCC(Mel Frequency Cepstral Coefficient) 특징 기반의 discrete HMM을 사용하고, 스켈레톤의 경우 가장 관련성이 높은 4개의 관절의 3차원 위치 및 DTW (Dynamic Time Warping) 거리를 기반으로 하여 kNN (k Nearest Neighbor) 방법을 적용한다. 최종적으로 두 분류기의 결과를 선형 결합하여 분류를 하는데, 실험결과를 통해 multi-modal 정보의 사용이 개별정보를 사용할 경우보다 인식결과를 향상시킴을 보여주었다.

2. 시공간적 접근법

먼저 참고문헌 [16]에서는 주어진 일련의 깊이 영상들을 4차원 시공간체적(4D space-time volume)으로 간주하고 그 안에서 사람에 해당하는 3차원 깊이 정보들의 분포를 묘사하는 STOP(Space-Time Occupancy Patterns) 특징을 제안하였다. 비슷하게 참고문헌 [17]에서는 4차원 시공간체적 안에서 다른 위치 및 다른 크기를 가지는 4차원 부속체적(4D sub-volume)들을 무작위로 샘플링하여 이를 활용하는 ROP(Random Occupancy Patterns) 특징을 제안하였다. 전체적(holistic) 특징인 STOP에 비해 ROP는 국소적(local) 특징이라고 볼 수 있다. 또한 참고문헌 [17]은 동작 분류를 위해 추가적으로 sparse coding을 적용하였으며, SVM(Support Vector Machine)을 사용하였다. 참고문헌 [18]은 세 직교(orthogonal) 평면으로 투영된 깊이 영상의 움직임 정보를 나타내는 이진(binary) 영상들을 DMM(Depth Motion Map)이라고 정의하였다. 이로부터 HOG를 계산하여 DMM-HOG 특징 벡터를 제안하였는데, 현재까지 알려진 바로는 MSR Action3D dataset에 대해 가장 높은 분류 성능을 보인다. 참고문헌 [19]의 경우 스켈레톤으로부터 자세, 움직임, 그리고 옵셋(offset) 정보를 계산하고 PCA를 적용하여 Eigen Joints 특징 벡터를 생성한다. 이 논문은 특징적으로 동작-동작(action-action) 거리가 아닌 동작-클래스(action-class) 거리를 계산하는 Naïve-Bayes-Nearest-Neighbor classifier를 사용한다. 마지막으로 참고문헌 [20]에서는 기존의 오프라인(offline)이 아닌 온라인(online) 동작인식 문제를 다루었는데, 이를 위해 각각의 동작이 인식되어야 하는 시점을 나타내는 action point라는 개념을 제안하였다. 효율적인 온라인 동작인식을 위해 무작위 결

정 트리(randomized decision trees) 방법을 제안하였고, 실험을 통해 대기 시간(latency)과 인식 정확도를 나타내는 F-score 사이에 트레이드오프(tradeoff) 관계가 있음을 입증하였다.

Ⅳ. 손 동작인식 기술

손은 사용자가 가장 손쉽게 사용할 수 있으며 자유도가 높아 다양한 표현을 할 수 있는 매개체이다. 이러한 손을 추적하며 동작을 분석하기 위해 과거에는 컬러센서를 통해 촬영된 영상에서 손을 찾아 분할하고 분할된 영상에서 손의 자세를 인식하는 등의 접근을 해 왔다. 이러한 컬러영상을 활용하던 방법들을 너머, 최근에는 거리영상을 얻을 수 있는 장비들이 상용화 됨에 따라 거리 정보를 이용하여 사용자 손을 분석하고 동작을 추정하는 연구들이 활발하게 진행되고 있다.

본 장에서는 이러한 손 동작을 위한 기술들을 손 탐색(hand detection), 손 자세 추정(hand posture estimation), 그리고 손 동작 분류(hand gesture classification)의 크게 세가지 경우로 나누고 각각의 최근 기술동향에 대해서 기술하겠다[21]-[23].

1. 손 탐색

손에 대한 정보를 얻기 위해서는 손이 실제 화면 상에 존재하는지 여부를 먼저 탐색해야 한다. 가장 기본적인 방법은 입력 컬러영상에서 미리 모델링된 피부색(skin-color) 정보로 손을 탐색하는 것이다[24]-[27]. 이 경우 기존에는 배경이 복잡하거나 피부색과 유사한 객체가 배경에 존재할 경우에 알고리즘의 성능이 저하되는 경향이 있었으나, 추가로 거리정보를 참고하면 손 부분을 배경과 용이하게 분리할 수 있다는 장점이 있다.

다른 방법으로 손이 가지는 특징들을 영상으로부터 분석하여 손을 탐색하는 방법들이 있다. 참고문헌 [28]은 미리 촬영된 손 데이터들의 윤곽 정보를 참고하여 화면에서 손을 탐색하고, 참고문헌 [21]에서는 손 끝과 같이 뾰족한 특징(tip)을 가지는 부분과 손가락과 같이 관형태의 특징(pipe)을 가지는 영역을 영상으로부터 추출한 다음 이를 활용하여 손을 탐색하고 자세를 추정한다. 참고문헌 [29]의 경우에는 거리영상에서 특정 구간을 추출하여 손이라 가정하고 추출된 영역에서 손바닥의 중앙점을 정의한다. 그 후 이로부터 최대의 3차원 geodesic distance를 가지는 위치를 탐색하여 이를 손가락 끝(fingertip)으로 인식하고 이를 추적한다. 최근 동작인식 인터페이스에서는 손을 포함하는 사용자 골격(skeleton)을 3차원 위치정보로 제공하는데 이를 활용하여 얻어진 손 위치 정보를 가지고 자세 추정 등에 활용하기도 한다[30].

한편, 손 탐색 알고리즘에 초점을 맞추지 않은 여러 연구들에서는 '사용자 앞에 별 다른 객체가 존재하지 않을 것이고, 거리영상에서 최소의 점이 손일 것이다'라는 가정을 활용하여 거리영상의 최소점에서 특정 구간을 정의하여 손이라 간주하기도 한다[31]-[33].

2. 손 자세 추정

손 자세(hand posture)는 연속적인 동작이 아닌 단일 영상에서 사용자가 취한 손 형태를 의미한다. 본 절에서는 손 자세 추정동향을 기존의 논문들에서 소개된 분류체계인 형태 기반 접근방법(shape-based approach)과3차원 모델 기반 접근방법(3D model-based approach)으로 나누어 소개한다[22][23].

가. 형태 기반 접근방법

형태 기반 접근방법을 가지는 손 자세 추정 알고리즘은 컬러영상 혹은 거리영상 등과 같은 입력영상으로부터 손에 해당하는 부분의 특징들을 추출하여 현재 손 자세를 구분하는 방법이다. 참고문헌 [32], [34]는 분할된

손에서 원 형태로 추출한 샘플의 패턴을 분석하여 사용자의 현재 손 자세를 추정한다. 손가락보다 다소 두꺼운 손목의 위치를 포함하는 패턴이 나오기 때문에 같은 손 자세일 경우 사용자 손의 롤(roll) 회전에 영향을 받지 않고 자세를 판별해 낼 수 있다는 장점이 있다. 한편 분할된 손의 윤곽선에서 특징을 추출하여 손 자세 판별에 활용하기도 하는데 윤곽선에서 급격하게 꺾이는 곡선의 개수를 분석하여 사용자가 몇 개의 손가락을 펴고 있는지를 분석한 연구가 있으며[31], 분할된 손의 윤곽선을 구한 다음 미리 학습된 데이터베이스에 있는 손의 윤곽선과 차이를 계산하여 손 자세를 추정하기도 한다[30]. 특정 자세의 손 형태정보를 미리 수집하고 학습하여 판별하는 방법을 기반으로 다양한 연구들이 진행되어 왔다. 참고문헌 [24]는 가상환경에서 손을 합성하여 윤곽과 거리정보를 뽑아내어 활용하며, 참고문헌 [35]는 컬러장갑을 끼고 촬영한 학습데이터를 활용하여 입력영상에서 2D 윤곽, 3차원 포인트 클라우드(3D point cloud)에서 추출한 표면정보, 그리고 손의 위치를 특징으로 입력영상의 손 자세를 추정한다. 최근에는 참고문헌 [36]처럼 거리 영상센서에서 나오는 3차원 포인트 클라우드 정보에서 3차원 접평면(tangent plane)의 법선 벡터(normal vector)를 추출하는 등 3차원 점 데이터로부터 3차원 정보를 재가공해서 손 자세 추정의 특징으로 사용하는 접근이 이루어지고 있다. 이런 접근들은 거리정보의 유용성을 보여준다고 할 수 있다.

나. 3차원 모델 기반 접근방법

3차원 모델 기반 접근방법은 입력된 영상에서 촬영된 사용자의 실제 손과 3차원으로 모델링 된 손의 불일치도(discrepancy)를 최소화시켜 맵핑시키는 방법으로 손 자세의 연속적인 추적을 목표로 하는 접근방법이다. 참고문헌 [25]는 손을 해부학적으로 분석하여 27개의 변수 형태로 표현하여 모델링한다. 그리고 가상 손 모델과 실제 손의 차이를 구하기 위해서 거리영상과 컬러정보 등을 활용하여 불일치도를 정의하고 이를 최소화시키기 위해 최적화 알고리즘 중 하나인 Particle Swarm Optimization(PSO)을 사용하여 한 손의 자세를 추정한다. 나아가 참고문헌 [37]에서는 양 손의 자세를 최적화 알고리즘으로 구하면서도 '양손 깍지끼기'와 같이 양 손 사이의 긴밀한 상호작용의 경우도 지속적으로 추적할 수 있는 기술을 소개했고, 최근에는 손과 컵, 그릇과 같은 실제 객체와의 상호작용(예. 손으로 컵을 잡고 들어 옮기기)들도 추적하여 손뿐만 아니라 손과 객체 사이의 상호작용 또한 추적하고 인식할 수 있음을 보였다[38].

3. 손 동작 분류

손 특징을 추출하는 알고리즘이 다양하듯 손 동작을 분류하는 알고리즘도 다양한 방법들이 존재한다. 참고문헌 [26], [36]에서는 대표적인 기계학습(machine learning) 알고리즘 중의 하나인 SVM을 활용한다. 추출된 특징들을 기준으로 학습시킨 SVM을 활용하여 입력영상의 특징과 가장 가까운 데이터셋을 찾고 이를 판별결과로 출력한다. 여러 가지 분류기들의 성능을 비교하여 결과를 보여준 논문인 참고문헌 [32]에서는 세 가지의 분류기들, 즉 k-d tree based K-means clustering, Bayesian plug-in classifier, 그리고 nearest neighbor classifier의 성능을 비교했으며, 결론적으로 해당 논문의 실험 환경에서는 nearest neighbor 방법이 가장 좋은 성능을 보였다. 손 동작 분류의 정확도를 높이기 위해서 하나 이상의 분류기를 함께 사용하는 연구들도 존재한다[27][28], [39][40]. 그 들 중 특히 [28], [39]의 경우에는 다수의 분류기를 계층적으로 배치시켜 적용함으로써 분류 성능을 향상시켰다. 또 다른 연구로는 손 위치의 연속적인 변화 패턴이 하나의 동작이 되기 때문에 동적 패턴 분석에 많이 사용되었던 HMM이나[41][42], 특정 동작이 이루어지는 속도나 시간에 영향을 받지 않는 것을 특

징으로 가지는 DTW[24]를 사용해서 연속적인 동작을 구분하는 연구들도 있다.

한편, 손 동작 분류 알고리즘들의 성능을 평가하기 위하여 필요한 표준 데이터셋이 갖추어지지 않는 실정이지만[23], 많은 연구들에서 손 동작 알고리즘의 성능을 평가하고자 할 때 수화(手話, sign language)를 얼마나 정확하게 구별하는지를 확인하여 평가한다[26], [36], [40]. 수화는 다양한 손 모양을 포함하고 있는데, 2차원 윤곽으로는 다소 구별하기 힘든 동작들도 포함되어 있어 거리영상 기반 손 동작 알고리즘을 평가하기에 적합하다. 하지만 이 역시 나라별로 수화 표현이 다르기 때문에 손 동작인식 기술들의 객관적인 평가를 위해서는 표준 데이터셋이 필요한 실정이다.

V. 결론

동작인식은 영상을 통해 대상 사람의 움직임을 분석하고 어떠한 자세 혹은 동작을 취했는지 판단하는 기술이다. 이는 최근 큰 주목을 받으며 다양하게 적용되고 있는 동작 기반 사용자 인터페이스의 기반이 되는 기술이며, 영상 기반의 보안, 영유아나 노인과 같은 취약자에 대한 모니터링, 그리고 장기적으로는 사람의 행동을 이해하고 적절한 반응을 취해야 하는 지능 로봇 등을 위한 핵심기술이라고 할 수 있다. 본고에서는 이러한 동작인식 기술 중에서도 거리영상 기반의 동작인식에 대한 최신 연구동향을 살펴보았다.

거리영상 기반 동작인식 기술은 현재 그 성능 및 속도에 있어서 만족할만한 수준에 도달하여 실제적인 적용이 이루어지고 있는 상황이다. 하지만 매우 간단한 배경에서 가리워짐(occlusion)이 그다지 크지 않은 동작을 대상으로 하고 있는 경우가 대부분이기 때문에 이러한 점에서의 기술개발 및 성능향상이 필요하다고 보여진다. 또한 센서에서 대상 사람까지의 거리가 현재 최대 1~4m 정도에 머물고 있는 상황인데, 거리 센서 자체의 센싱 범위 확장에 대한 연구가 필요한 동시에 원거리 저해상도 영상에서의 동작인식에 대한 연구도 이루어져야 할 것이다. 그리고 현재에는 대상 사람의 수가 1~2 명 정도로 제한적인 경우가 많으며, 이를 증가시키기 위해 기존 기법의 계산량을 줄이는 연구 또한 필요하다고 생각된다. 또한 이렇게 사람의 수가 늘어나는 경우 동일한 사람을 시간축 상에서 연속적으로 인식하기 위해 다수 사람 추적(human tracking)기술이 적용되어야 할 것이다. 마지막으로 인식대상인 동작의 복잡도를 고려해 볼 수 있는데, 동작은 그 복잡도에 따라 사람 동작의 가장 단순한 단위로 간주되는 gesture, 몇 가지 gesture들의 결합을 통해 이루어지는 action, 그리고 사람과 사람 사이 혹은 사람과 물체 사이의 상호 작용을 의미하는 interaction의 세 가지 경우로 대략 나눌 수 있다. 현재 주로 gesture 혹은 간단한 action에 대한 인식에 머물러 있는 기술수준을 더욱 복합적인 action 및 interaction 인식수준으로 끌어 올리는 것이 동작인식의 최종 목표가 될 것이다.

참고문헌

[1] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," *IEEE Conf. Comput. Vision Pattern Recognition(CVPR)*, 2011.

[2] G. Sansoni, M. Trebeschi, and F. Docchio, "State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation," *Sensors*, vol. 9, no. 1, pp. 568-601, 2009.

[3] P. Hariharan, *Optical interferometry*. Elsevier, 2003.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International J. Comput. Vision,* vol. 47, no. 1-3, 2002, pp. 7-42.

[5] S. M. Seitz et al., "A comparison and evaluation of multi-view stereo reconstruction algorithms," *IEEE Conf. Comput. Vision Pattern Recognition*, vol. 1, 2006, pp. 519-528.

[6] J. Salvi et al., "A state of the art in structured light

patterns for surface profilometry," *Pattern recognition*, vol. 43, no. 8, 2010, pp. 2666–2680.

[7] S. Lee, O. Choi, and R. Horaud, Time-of-flight cameras: *Principles, methods and applications. Springer*, 2013.

[8] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (tof) cameras: A survey," *Sensors,* vol. 11, no. 9, 2011, pp. 1917–1926.

[9] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," *Comput. Vision Pattern Recognition Workshops(CVPRW),* 2010 *IEEE Comput. Soci. Conf.,* 2010, pp. 9–14.

[10] J. Wang et al., "Mining actionlet ensemble for action recognition with depth cameras," *IEEE Conf. Comput. Vision Pattern Recognition(CVPR)*, 2012, pp. 1290–1297.

[11] S. Escalera et al., "Multi-modal gesture recognition challenge 2013: Dataset and results," *ChaLearn Multi-modal Gesture Recognition Grand Challenge Workshop, 15th ACM International Conf. Multimodal Interaction,* 2013.

[12] A. Jalal et al., "Recognition of human home activities via depth silhouettes and r transformation for smart homes," *Indoor Built Environment,* vol. 21, no. 1, 2012, pp. 184–190.

[13] L. Xia, C.-C.Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," *IEEE Conf. Comput. Vision Pattern Recognition Workshops(CVPRW),* 2012, pp. 20–27.

[14] J. Sung et al., "Unstructured human activity detection from rgbd images," *IEEE International Conf. Robot. Autom(ICRA).*, 2012, pp. 842–849.

[15] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," *Proc. 15th ACM International Conf. Multimodal Interaction*, 2013, pp. 453–460.

[16] A. W. Vieira et al., "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," *Progress Pattern Recognition, Image Anal. Comput. Vision, Applications*. Springer, 2012, pp. 252–259.

[17] J. Wang et al., "Robust 3d action recognition with random occupancy patterns," *ECCV*. Springer, 2012, pp. 872–885.

[18] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," *Proc. 20th ACM international Conf. Multimedia*, 2012, pp. 1057–1060.

[19] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," *IEEE Conf. Comput. Vision Pattern Recognition Workshops(CVPRW)*, 2012, pp. 14–19.

[20] S. Nowozin and J. Shotton, "Action points: A representation for low-latency online human action recognition," Technical report, Tech. Rep., 2012.

[21] G. Hackenberg, R. McCall, and W. Broll, "Lightweight palm and finger tracking for real-time 3d gesture control," *Virtual Reality Conf(VR).*, 2011, pp. 19–26.

[22] J. Han et al., "Enhanced computer vision with microsoftkinect sensor: A review," *IEEE Trans. Cybern.*, 2013.

[23] M. Ye et al., "A survey on human motion analysis from depth data," *Time-of-Flight and Depth Imaging. Sensors, Algorithms, Appl.* Springer, 2013,pp. 149–187.

[24] P. Doliotis et al., "Comparing gesture recognition accuracy using color and depth information," *Proc. 4th International Conf. Pervasive Technol. Related Assistive Environments*, 2011, pp. 20:1–20:7.

[25] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," *BMVC*, 2011, pp. 1–11.

[26] H. Takimoto et al., "Classification of hand postures based on 3d vision model for human-robot interaction," *RO-MAN, 2010 IEEE*, 2010, pp. 292–297.

[27] M. Van den Bergh and L. Van Gool, "Combining rgb and tof cameras for real-time 3d hand gesture interaction," *IEEE Workshop Appl. Comput. Vision (WACV)*, 2011, pp. 66–72.

[28] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," *Sixth IEEE International Conf. Autom. Face Gesture Recognition*, 2004, pp. 889–894.

[29] H. Liang, J. Yuan, and D. Thalmann, "3d fingertip and palm tracking in depth image sequences," *Proc. 20th ACM International Conf. Multimedia,* 2012, pp. 785–788.

[30] M. Caputo et al., "3d hand gesture recognition based on sensor fusion of commodity hardware," *Mensch & Comput.* 2012, pp. 293–302.

[31] H. Lahamy and D. Lichti, "Real-time hand gesture recognition using range cameras," *Proc. Canadian Geomatics Conf.*, 2010.

[32] S. Soutschek et al., "3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras," *IEEE Conf. Comput. Vision Pattern Recognition Workshops(CVPRW)*. 2008, pp. 1–6.

[33] R. Tara, P. Santosa, and T. Adji, "Hand segmentation from depth image using anthropometric approach in natural interface development," 2012.

[34] S. E. Ghobadi et al., "Real time hand based robot control using multimodal images," *IAENG International J. Comput. Sci.*, vol. 35, no. 4, 2008, pp. 500–505.

[35] Y. Yao and Y. Fu, "Real-time hand pose estimation from rgb-d sensor," *IEEE International Conf. Multimedia Expo(ICME)*, 2012, pp. 705–710.

[36] C. Zhang, X. Yang, and Y. Tian, "Histogram of 3d facets: A characteristic descriptor for hand gesture recognition," *International Conf. Autom. Face Gesture Recognition*, 2013.

[37] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Tracking the articulated motion of two strongly interacting hands," *IEEE Conf. Comput. Vision Pattern Recognition(CVPR)*, 2012, pp. 1862–1869.

[38] N. Kyriazis and A. Argyros, "Physically plausible 3d scene tracking: The single actor hypothesis," *Proc. 2013 IEEE Conf. Comput. Vision Pattern Recognition*, 2013, pp. 9–16.

[39] Y. Li, "Hand gesture recognition using kinect," *IEEE 3rd International Conf. Softw. Eng. Service Sci (ICSESS)*., 2012, pp. 196–199.

[40] D. Uebersax et al., "Real-time sign language letter and word recognition from depth data," *IEEE International Conf. Comput. Vision Workshops(ICCV Workshops)*, 2011, pp. 383–390.

[41] C. Keskin, A. Erkan, and L. Akarun, "Real time hand tracking and 3d gesture recognition for interactive interfaces using hmm," *ICANN/ICONIPP*, 2003, pp. 26–29.

[42] B. Wang, Z. Chen, and J. Chen, "Gesture recognition by using kinect skeleton tracking system," *5th International Conf. International Conf. Intelligent Human-Mach. Syst. Cybern(IHMSC)*, vol. 1, 2013, pp. 418–422.