논문 2006-43CI-3-5

강화 학습을 이용한 비전 기반의 강인한 손 모양 인식에 대한 연구

(A Study on Vision-based Robust Hand-Posture Recognition Using Reinforcement Learning)

장 효 영*, 변 증 남**

(Hyoyoung Jang and Zeungnam Bien)

요 약

본 논문에서는 비전 기술에 기반을 둔 손 모양 인식 시스템의 성능 향상을 위하여 강화학습에 의한 손 모양 인식 방법을 제안한다. 비전 센서에 기반을 둔 손 모양 인식은 손의 높은 자유도로 인한 자체 겹침 (self-occlusion) 현상과 관찰 방향 변화에 따른 입력 영상의 다양함으로 인식에 어려움이 따른다. 따라서 비전 기반 손 모양 인식의 경우, 카메라와 손 간의 상대적인 각도에 제한을 두거나 여러 대의 카메라를 배치하는 것이 일반적이다. 그러나 카메라와 손 간의 상대적 각도에 제한을 두는 경우에는 사용자의 움직임에 제약이 따르게 되며, 여러 대의 카메라를 사용할 경우에도 각 입력된 영상에 대한 인식 결과를 최종 인식 결과에 반영하는 방식에 대하여 추가적인 고려를 해야 한다. 본 논문에서는 비전 기반 손 모양 인식의 이러한 문제점을 개선하기 위하여 인식 과정에서 사용되는 특징을 손 구조적인 각도 정보와 손 윤곽선 정보로 나누고 강화학습을 통하여 각 특징간의 연관성을 정의하는 방식을 제안한다. 또한 제안된 방법을 세 대의 카메라를 이용한 손 모양 인식 시스템에 적용하여 유용성을 검증한다.

Abstract

This paper proposes a hand-posture recognition method using reinforcement learning for the performance improvement of vision-based hand-posture recognition. The difficulties in vision-based hand-posture recognition lie in viewing direction dependency and self-occlusion problem due to the high degree-of-freedom of human hand. General approaches to deal with these problems include multiple camera approach and methods of limiting the relative angle between cameras and the user's hand. In the case of using multiple cameras, however, fusion techniques to induce the final decision should be considered. Limiting the angle of user's hand restricts the user's freedom. The proposed method combines angular features and appearance features to describe hand-postures by a two-layered data structure and reinforcement learning. The validity of the proposed method is evaluated by appling it to the hand-posture recognition system using three cameras.

Keywords: *hand-posture recognition, reinforcement learning, view-invariant*

I. 서 론

최근 '유비쿼터스 컴퓨팅(ubiquitous computing)' 에 대한 관심이 높아지고 있다. 이는 컴퓨터, 통신, 그리고 이를 이용하는 인간이 조화롭게 연결될 수 있는 환경을 의미한다. 유비쿼터스 컴퓨팅의 개념을 제시한 마크 와이저(Mark Weiser)는 유비쿼터스 컴퓨팅의 주요 특징 중 하나로서 인간화 인터페이스의 중요성을 강조하였다.[1] 컴퓨터 환경의 발전에는 그에 적합한 사용자 인터페이스(user interface; UI)의 변혁이 수반된다. 초기의 텍스트 위주의 컴퓨터 환경에서는 키보드가 주된 UI이었으나, 이후 GUI (graphic user interface) 환경으로의 변화는 키보드에 더하여 마우스를 주요한 사용자 인터페이스로 부각시켰다. 마찬가지로 근래의 사용자를 중심으로 하는 기술 개발의 경향은 지능을 가진 자동 인

* 학생회원, ** 평생회원, KAIST 전자전산학과
(Department of Electrical Engineering and Computer Science, KAIST)
※ 본 연구는 과학기술부/한국과학재단 우수연구센터 육성사업의 지원으로 수행되었음 (R11-1999-008)
접수일자: 2006년3월16일, 수정완료일: 2006년5월8일

터페이스로의 전환을 촉구하고 있다. 손 제스처는 이에 대한 하나의 안으로 부각되고 있다.

손 제스처는 일상생활에서 사람 간 언어적인 특질을 지닌 독립적인 대화 수단으로 다양하게 활용된다. 물건을 지시하는 것과 같은 단순한 것에서부터 수화와 같이 대화를 목적으로 하는 복잡한 것까지 그 활용이 다양하다. 대표적인 의사 전달 수단의 하나인 음성의 보조 수단으로 사용할 뿐만 아니라, 소음이 심한 공사 현장이나 수중 등 음성을 통한 의사 전달이 불가능 한 경우 음성을 대체하는 수단으로써 사용되기도 한다. 손 제스처는 특히 방향이나 사물의 지시 등과 같이 음성 명령으로는 표현에 제약이 따르는 공간적인 정보의 표현에도 용이하다. 이와 같은 친숙함과 음성 대체 수단으로서의 기능성, 그리고 공간 표현에 대한 적합성으로 인하여 가상현실(virtual reality)과 수화 인식 분야를 포함한 여러 분야에서 손 제스처 인식에 기반을 둔 사용자 인터페이스에 대한 연구가 진행되어 왔다.[2][3]

본 논문은 손 제스처를 구성하는 요소* 중, 손 모양을 비전 센서로 인식하는 경우의 성능을 향상시킬 수 있는 방법으로서 OSS-Net(Object Shape and Structure Network)를 제안한다.

논문의 구성은 다음과 같다. Ⅱ장에서는 기존 연구의 접근 방식에 대해 간략하게 소개하고 문제점을 밝힌다. Ⅲ장에서는 제안하는 방식의 데이터 구조를 기술하며, Ⅳ장에서는 OSS-Net을 이용한 손 모양 인식 과정을, 그리고 Ⅴ장에서는 OSS-Net의 구축과 학습 과정에 대하여 설명한다. Ⅵ장에서 3대의 카메라를 이용한 손 모양 인식 시스템에 대한 적용 예를 보이고 Ⅶ장에서 논문을 마친다.

Ⅱ. 기존 연구 및 문제점

손 모양 인식에서 가장 큰 화두는 사람 손의 높은 자유도와 관측 방향에 따른 다양한 형태 발생에 대한 것이다. 사람의 손은 27개의 뼈로 이루어져 움직임에 따른 형태 변형이 크다.[4] 또한 동일한 손 모양이라 하더라도 관찰 방향이 바뀌게 되면 영상을 통해 취득한 손 모양의 외관(appearance)이 크게 변하게 된다. 대개의 경우, 손의 형태적인 변형과 관찰 방향 변화는 복합적으로 발생하며, 이는 비전을 통한 손 모양 인식을 어렵게 하는 요인이 된다.

손 모양 인식의 과정은 크게 두 단계로 나뉜다. 첫째가 손을 표현하는 특정 모델(feature-based model)을 두어 손의 특징 값(feature)을 표출하는 부분이고 둘째가 이 특징 값을 기 정의된 손 모양에 대한 사전 정보와 정합하는 단계이다.

손 모델을 정의하는 방식으로는 크게 2차원 형상에 기반을 두어 손 모델로 정의하는 방식(2-dimensional appearance-based approach)과 3차원 손 모델로 정의하는 방식(3-dimensional model-based approach)이 있다.

3차원 손 모델을 이용하는 방식은 3차원 캐드(CAD) 모델을 이용한다.[6][7] 이 방식은 손 모양의 높은 형태 변형가능성으로 인해 발생하는 문제를 해결하는 데에 주안점을 둔다. 입력 영상에 대해 국부 영상 특징(local image feature)을 추출한 후 이것을 다시 3차원적인 형태 모델이나 손의 구조를 모사한 뼈대 모델(skeleton model)과 비교하여 손 모양을 인식한다. 그러나 기존 모델과의 비교를 위해서는 입력된 손 모양으로부터도 역시 저장된 것과 같은 형식의 3차원 모델을 추출해야 하므로 이 과정에서 많은 계산 부하가 발생한다. 더욱이 손은 움직임에 의해 형태가 변화하는 대상이므로 영상으로부터 정확한 특징 점을 찾기가 어렵다. 따라서 별도의 표식을 손에 부착하거나, 사전에 손바닥과 카메라의 상대적인 방향을 정하는 등의 방법으로 움직임에 제한을 두기도 하나, 이는 사용자 편의성 관점에서 바람직하지 않다.

2차원 형상에 기반을 둔 손 모양 인식을 위해서는 다양한 조건에서 보이는 손 모양에 대한 여러 장의 영상으로 만든 모델을 이용한다.[8][9] 이 방식은 인간의 시각 체계(human visual system)가 3차원 모델에 기반을 두고 물체를 인식하기보다는 2차원 형상을 기반으로 물체를 인식한다는 최근의 연구 결과에 근거를 둔 것이다.[10] 2차원 형상 모델을 이용할 경우에는 입력 영상과 기 정의된 모델을 직접 비교하기 때문에, 영상에서 특징 점을 찾아 이로부터 3차원 모델을 구축하여 기존에 저장된 3차원 손 모델과 비교를 행하는 방식보다는 자체 가림 현상에 비교적 강하다. 그러나 방향에 강인한 인식 성능을 얻기 위해서는 다양한 방향에서 관측한 손

* 현존하는 제스처의 여러 형태 중 수화는 명확한 언어 체계를 가졌으며, 다양한 의미의 손짓에 대하여 명확하게 정의되어 있어, 수화의 분석 접근 방식은 손 제스처의 분석에도 동일하게 적용할 수 있다. 수화는 수화소라고 불리는 4개의 기본 요소에 의해 분석할 수 있는데, 이 4가지 기본 요소는 각각 손 운동(Motion-direction), 손 모양(Posture), 손위치(Location), 손 방향(Orientation)이다.[5]

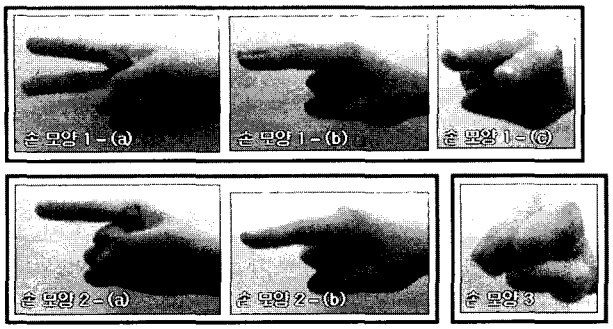

그림 1. 관측 방향에 따른 서로 다른 손 모양의 형상 구별 오류 발생 예

Fig. 1. Similar hand-appearance from different hand-posture examples under viewing-direction variation.

영상을 이용하여 모델을 정의해야 하는데, 한 손 모양에 대해 관측 방향의 수가 많아질수록 해당 손 모양에 대한 방향 강인성에 반비례하여 서로 다른 손 모양 간의 구분성이 낮아지게 된다. 그림 1에 관측 방향을 달리하여 취득한 손모양의 예를 보였다. 서로 다른 손 모양임에도 불구하고 손 영역의 윤곽선을 보았을 때 손 모양 1-(b)는 손 모양 2-(b)와 유사하며, 손 모양 1-(c)는 손 모양 3과 유사함을 확인할 수 있다.

관측 방향에 강인한 인식 성능을 얻기 위한 가장 보편적인 방법은 여러 대의 카메라를 통해 여러 방향에서의 영상을 취득하고 각각의 손 영상을 인식하여 최종결과에 반영하는 것이다. 이 경우, 여러 영상으로부터 얻은 각각의 인식결과의 조합으로부터 가장 높은 가능성을 갖는 최종 인식 결과를 이끌어내기 위하여 추가적인 방법이 필요하게 된다. 그러나 여러 대의 카메라를 이용하는 손 모양을 인식하는 기존 연구는 각 카메라로부터 취득한 영상 인식 결과로부터 최종 인식 결과를 도출해내는 과정에서 체계적인 방법론을 제시하지 못하고 있다. 응용 예에 따라 취득한 손 영역의 크기 또는 특징점의 개수, 이전 영상과의 편차, 움직임의 크기 등을 이용하여 실험에 의해 가정을 두어 최종 결과에 반영할 뿐이다.

III. Object Shape and Structure Network (OSS-Net)

'Object Shape and Structure Network (OSS-Net)' 모델은 본 논문을 통해 제안하는 손 모양 인식 시스템의 근간을 이루는 손 모양 데이터베이스 구조 및 인식 방법을 의미한다. 이 구조는 인식의 대상이 되는 개체의 2차원적인 형태 특성 뿐 아니라 3차원적인 구조 특징 또한 표현하도록 구성되어 있다.

OSS-Net 모델은 크게 2차원 형상 기반 특징 층(2-dimensional appearance-based feature layer)과 3차원 구조 기반 특징 층(3-dimensional structure-based feature layer)의 이층 구조로 이루어진다. 각 층은 주어진 손 모양을 기술하는 여러 노드(node)들로 구성된다. 즉, 2차원 형상 기반 특징 층에서 각 노드는 해당 손 모양에 대한 2차원 형상 특징을 나타내는 값들로 이루어진 벡터이며, 3차원 구조 기반 특징 층의 노드는 해당 손 모양에 대해 손의 구조적인 특징을 나타내는 값들로 이루어진 벡터이다. 그림 2에 OSS-Net의 개념을 간략하게 도시하였다.

OSS-Net은 다음과 같이 정의된다.

정의 1. OSS-Net, *oss_net* = *(S, P, L1, L2, C)*

S : 패턴 집합

P : 네트워크 파라미터

L1 : 2차원 형상 기반 특징 층
(2차원 형상 특징을 나타내는 노드의 집합)

L2 : 3차원 구조 기반 특징 층
(3차원 구조 특징을 나타내는 노드의 집합)

C : *L1*, *L2* 간 상호 연결

OSS-Net은 패턴 집합(*S*)과 네트워크 파라미터(*P*), 2차원 형상 기반 특징 층(*L1*), 3차원 구조 기반 특징 층(*L2*), 그리고 층간 상호 연결(*C*)로 구성된다. 패턴 집합 S는 OSS-Net의 생성 및 테스트를 위한 데이터를 의미한다. 네트워크 파라미터 P는 OSS-Net 구조에서 인식을 위한 최소한의 경계치(threshold value)와 최대 반복

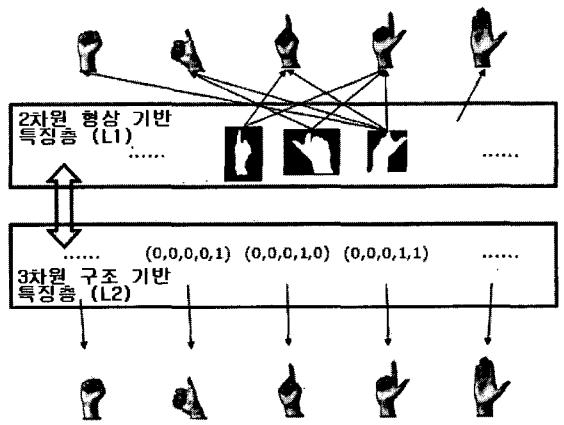

그림 2. OSS-Net 모델의 개념도

Fig. 2. Conceptual figure of OSS-Net Model.

회수 등의 변수를 나타내는 벡터이다.

다시 2차원 형상 기반 특징 층(*L1*)은 정의 2와 같이 정의된다.

정의 2. 2차원 형상 기반 특징 층, $L1 = (X, U)$

2차원 형상 특징 노드 집합, $X = \{\underline{x_1}, \underline{x_2}, \cdots, \underline{x_m}\}$,

m: 형상 특징 노드의 수,

$\underline{x_i} = (\underline{xf_i} : \underline{xh_i})$,

$\underline{xf_i}$: 2차원 형상 기반 특징 벡터,

$\underline{xh_i}$: 정의된 인식 대상 손 모양과의 연관성을 나타내는 p차원 벡터

(p : 인식 대상 손 모양 수)

형상 특징 노드 간 연결. $U = \{u_{ij}\}$, $i, j = 1, \cdots, m$,

$$u_{ij} = \begin{cases} 0 & \text{if } i = j \\ p_u(i,j) & otherwise, \end{cases}$$

$p_u(i,j)$: x_i와 x_j 간 연결 강도

마찬가지로, 3차원 구조 기반 특징 층(*L2*)은 정의 3과 같다.

정의 3. 3차원 구조 기반 특징 층, $L2 = (Y, V)$

3차원 구조 특징 노드 집합, $Y = \{\underline{y_1}, \underline{y_2}, \cdots, \underline{y_n}\}$,

n: 3차원 구조 특징 노드의 수,

$\underline{y_i} = (\underline{yf_i} : \underline{yh_i})$,

$\underline{yf_i}$: 3차원 구조 기반 특징 벡터,

$\underline{yh_i}$: 정의된 인식 대상 손 모양과의 연관성을 나타내는 p차원 벡터

(p : 인식 대상 손 모양 수)

구조 특징 노드 간 연결. $V = \{v_{ij}\}$, $i, j = 1, \cdots, n$,

$$v_{ij} = \begin{cases} 0 & \text{if } i = j \\ p_v(i,j) & otherwise, \end{cases}$$

$p_v(i,j)$: y_i와 y_j 간 연결 강도

마지막으로 층간 연결(*C*)의 정의는 정의 4와 같다.

정의 4. 층간 연결, $C = \{c_{ij}\}$,

$$c_{ij} = \begin{cases} 1 & \text{if } x_j \text{ is related to } y_i \\ 0 & \text{otherwise} \end{cases},$$

$i = 1, \cdots, n, \quad j = 1, \cdots, m$

OSS-Net에서는 각 특징 노드 간 연결에 의하여 각 특징 층을 정의한다. 특징 노드의 수는 고려 대상이 되는 손모양의 수에 의존적이며, 실제 구현에서는 일차적인 비교를 위한 템플릿의 수에 해당한다. 또 각 노드 간 연결은 연결 강도를 나타내는 실수 값으로 존재하게 되는데, 이는 손 모양 간 유사도를 대체한다.

Ⅳ. OSS-Net을 이용한 손 모양 인식

앞 절에서 정의된 것과 같은 OSS-Net 모델이 실제로 어떻게 생성되는가를 다루기에 앞서, 손 모양 인식 과정에서 어떤 식으로 동작하게 되는지를 살펴보는 것이 전체 인식 방법을 구성하는 세부 알고리즘과 그 필요성을 이해하는 데에 도움이 될 것이다. OSS-Net을 이용한 인식 과정을 그림 3에 나타내었다.

그림에 나타낸 바와 같이 OSS-Net을 이용한 인식 과정에서는 각 손 모양을 정의하는 데이터베이스 내 모델들과의 비교 및 탐색의 과정이 *L1*, *L2* 간 상대 층에 대한 결과 반영 및 처리 대상으로 하는 층의 전환으로 이루어진다. 즉, 데이터 처리의 흐름이 *L1*→(*L2*→*L1*)→(*L2*→*L1*)→ … 과 같이 이루어진다.

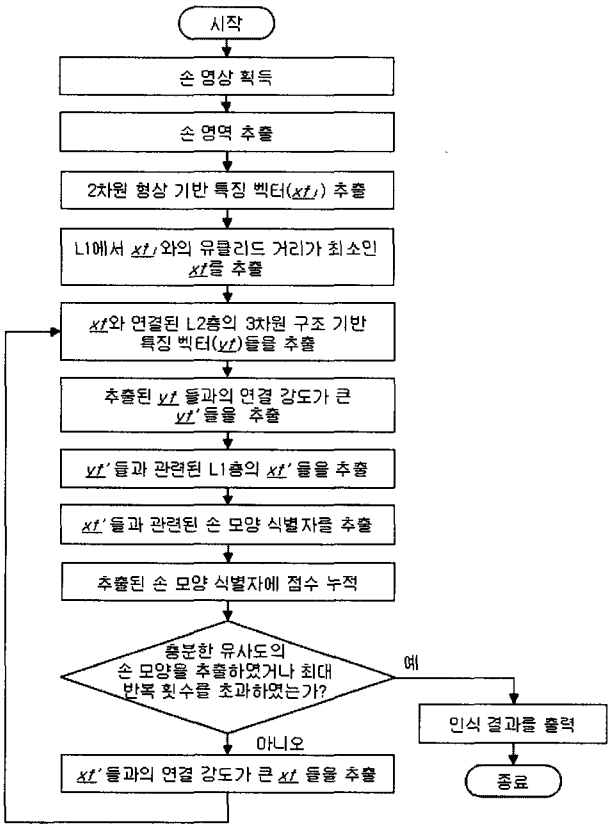

그림 3. OSS-Net 모델을 적용한 손 모양 인식 과정

Fig. 3. Hand-posture recognition process using OSS-Net.

L1 층과 *L2* 층을 구성하는 노드는 각기 2차원 형상 기반 특징과 3차원 구조 기반 특징을 표현한다. 층간 전환 과정을 통해 이 두 가지 특징 정보를 모두 고려하게 된다. 이와 같이 2차원 형상 기반 특징과 3차원 구조 기반 특징을 모두 사용하여 대상을 기술한 것은 손 모양 간 유사도를 결정함에 있어 두 가지의 기준이 존재할 수 있음을 전제하기 때문이다. 이것은 여러 방향에서 관찰한 물체에 대해 방향에 불변하게 인식을 하기 위해 반드시 필요하다.

OSS-Net을 통해 손 모양을 기술하고, 이 구조를 인식 과정에서 사용하기 위해서는 다음과 같은 사항이 필요함을 알 수 있다. V장에서 이에 대한 구체적인 사항을 다룬다.

특징 노드 생성 방법
노드 간 연결 생성 방법
유사도 결정 방식
노드 간 이동 결정 방식

이 중 노드 간 이동은 크게 두 가지로 나뉜다. 동일한 특징 층 내에서의 이동과 서로 다른 특징 층 간 이동이 그것이다. 동일한 특징 층 내에서의 이동은 2차원 특징 벡터를 기준으로 가장 유사한 손 모양을 찾거나, 3차원 특징 벡터를 기준으로 가장 유사한 손 모양을 찾는 과정이다. 서로 다른 특징 층 간 이동으로 xf 와 관련된 yf 를 추출하는 과정은 해당 손 영역 형상이 나올 수 있는 손 모양의 모든 경우를 찾는 과정이며, 또한 yf 와 관련된 xf 를 추출하는 과정은 특정 손 모양에 대해 관측 방향 변화에 따라 취득할 수 있는 손 형상들을 찾는 과정이다.

이와 같은 층간 탐색 과정의 반복을 통해 최종적으로 얻은 손 모양 인식 결과는 관측 방향 변화에 대해 강인한 성질을 갖게 된다.

V. OSS-Net 구축과 학습

OSS-Net 구축 과정은 크게 *L1*층과 *L2*층에 특징 노드를 추가하고 특징 노드 간 연결을 생성하는 과정으로 요약된다. 이를 위해서는 인식의 대상으로 하는 손 모양의 2차원 형상 특징 벡터와 함께 3차원 구조 특징 벡터 및 해당 손 모양에 대한 식별자(ID)가 주어져야 한다. 2차원 형상 특징 벡터와 3차원 구조 특징 벡터를 취득하고 기술하는 방법으로는 해당 응용 분야와 인식 대상으로 하는 손모양의 수 및 특성 등에 따라 여러 가지가 가능할 것이다.

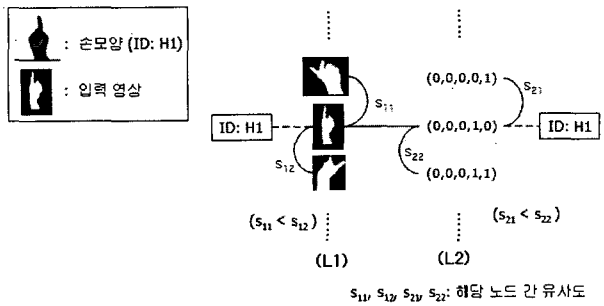

그림 4. OSS-Net 의 구축 예
Fig. 4. Example of OSS-Net.

이해를 돕기 위하여 그림 4에 구축된 OSS-Net의 간략한 형태를 보였다. 이 예에서는 2차원 형상 기반 특징으로, 입력 영상으로부터 추출한 손 영역 자체를 사용하며, 2차원 형상 기반 특징 층에서 각 노드 간의 유사도는 추출된 손 영역 영상 간 템플릿 정합(template matching)을 통한 템플릿 계수로 나타낸다.

템플릿 정합은 영상 속에 있는 모양과 크기, 방향이 알려진 객체를 검출하기 위한 방법으로, 직접적이고 효율적이기 때문에 응용 빈도가 높다. 찾고자 하는 형상으로 필터 마스크를 정의하며, 필터 마스크는 주어진 영상을 가로질러 움직이면서 컨벌루션(convolution)된다. 이 마스크가 찾고자 하는 영상 객체 위에 위치했을 때의 컨벌루션 결과는 최대 세기(intensity)를 갖는다. 결국, 찾고자 하는 객체의 영상 내 존재 여부와 위치는 컨벌루션 결과에서의 피크 검출에 의해 결정된다.

3차원 구조 기반 특징으로는 5개 손가락의 상태를 이용하며, 손가락을 폈을 경우에는 1로 그렇지 않을 경우는 0으로 나타내기로 한다. *L2* 층에서의 노드 간 유사도는 각 벡터간의 유클리드 거리(Euclidean distance)의 역수를 사용한다.

이와 같은 설정 하에서 그림 4와 같이 각 노드와 연결을 생성하고 손 영역 특징 간의 비교를 수행할 수 있다.

주어진 <2차원 형상 특징 벡터, 3차원 구조 특징 벡터, 손 모양 식별자> 집합에 대해 각 층의 특징 노드를 생성하는 과정은 그림 5와 같다.

먼저, 주어진 2차원 형상 특징 벡터를 OSS-Net의 *L1* 층을 구성하는 특징 노드의 특징 벡터 부분과 비교하여 가장 작은 유클리드 거리를 갖는 *L1*의 노드를 찾는다. 만약 이 노드와의 거리가 미리 정한 임계치보다 작을 경우에는 기존 노드가 입력된 벡터를 반영하도록

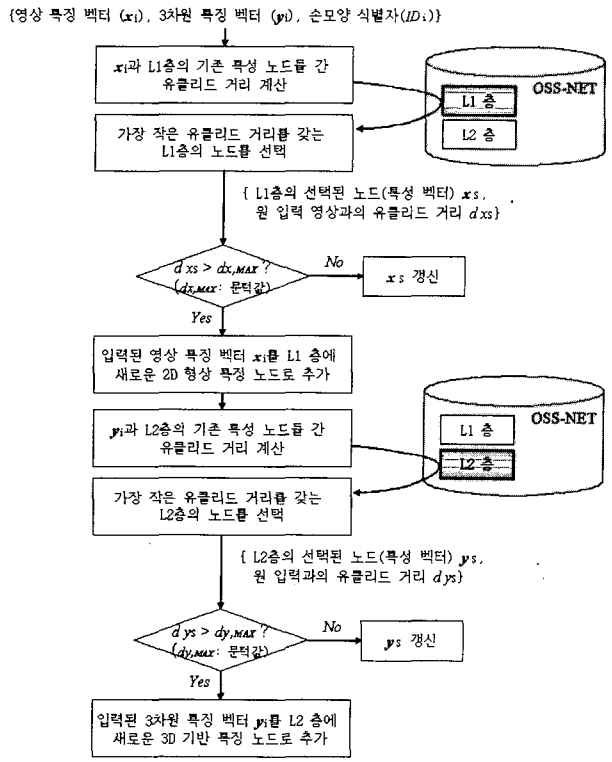

그림 5. 노드 생성 흐름도
Fig. 5. Flowchart of creating nodes.

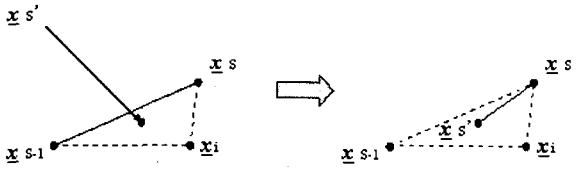

그림 6. 새로운 노드의 추가
Fig. 6. Creating a new node.

갱신한다. 입력된 형상 특징 벡터와 *L1* 층의 노드와의 거리가 모두 임계치보다 클 경우에는 주어진 형상 특징 벡터를 새로운 특징 노드로서 *L1* 층에 추가한다. 이어지는 *L2* 층과의 비교도 유사한 방법을 따른다. 주어진 3차원 구조 특징 벡터와 *L2* 층의 특징 노드들의 특징 벡터 부분을 비교하여 임계치 이상의 유클리드 거리를 가질 경우, 입력된 3차원 구조 특징 벡터가 *L2* 층에서의 새로운 특징 노드가 되고, 그렇지 않을 경우에는 기존 노드를 조정한다.

기존에 각 층에 존재하던 노드를 조정할 때의 방식은 *L1* 층과 *L2* 층에 대해 동일하다. 입력 특징 벡터와의 유클리드 거리가 가장 작은 기존 노드, $\underline{x_s}$와 동일 층에 속하는 기존 노드 중 $\underline{x_s}$와 가장 작은 유클리드 거리를 갖는 $\underline{x_{s-1}}$, 그리고 입력 특징 벡터 $\underline{x_i}$의 평균으로 식 1과 같이 새로운 노드의 특징 벡터 부분을 정의한다(그림 6).

$$\underline{x_s}' = \frac{\underline{x_s} + \underline{x_{s-1}} + \underline{x_i}}{3} \quad (1)$$

노드를 추가하거나 조정한 후에는 그에 따라 노드 간 연결 또한 조정해준다. 하나의 노드에 대하여 층 간 연결은 복수 개가 존재할 수 있다. 이는 동일한 손 모양에 대해 관찰 방향에 따라 외형이 다양하게 존재할 수 있음(또는 동일한 외형에 대해 실제로 손모양은 다를 수 있음)을 반영하는 것이다. 층 내 연결 또한 노드 간 이동의 정의 방식에 따라 여러 개가 존재할 수 있다.

층 간 연결 방법은 다음과 같다. 새로운 노드를 추가했을 경우에는 반대편 층의 동일한 식별자를 가진 모든 노드와 연결이 생성된다. 기존 노드를 갱신했을 경우에는 갱신된 노드가 기존 노드의 연결을 모두 계승한다. 층 내 연결은 기본적으로 같은 층에 존재하는 모든 노드에 대해 생성되며, 이후의 학습 과정을 통해 각 연결의 강도를 조절한다. OSS-Net 구조에서 해당 층 내 노드 간의 이동 확률을 결정하는 방식은 전체 인식 성능에 크게 영향을 미친다.

OSS-Net에서의 인식 과정은 주어진 입력 영상에서 취득한 손 영역과 가장 유사한 형태를 갖는 2차원 형상 특징 벡터에 해당하는 노드로부터 장기적인 관점에서 최적인 손 모양에 해당하는 대해 노드를 찾아가는 경로 생성 문제로 생각할 수 있다. 이에, 최적의 경로를 학습하기 위한 방법으로 강화 학습(reinforcement learning) 알고리즘을 적용하였다.

그림 7에 근시안적인 최적(myoptically optimal 또는 myoptimal) 손 모양 탐색 과정을 적용시켜 오류가 발생하는 경우의 예를 보였다. 그림 7에서는 2차원 특징량 간의 유사도를 템플릿 정합에 의해 결정하며, *L2* 층에서의 노드 간 이동시 유클리드 거리가 가장 작은 노드를 인접 노드로 선택한다. 이 때, 문제는 이러한 "보다 높은 유사도"를 정의하는 과정에서 비롯된다. 템플릿 정합 결과가 높은 것이 반드시 원 손 모양에 가까운 손 모양인 것은 아니며, 마찬가지로 유클리드 거리 비교를 통해 3차원 특징량을 비교한다고 했을 경우에 가장 작은 거리를 갖는 3차원 특징량이 2차원 영상 측면에서 보다 가까운 손 모양을 의미하지는 않는다. 다시 말하자면 손모양의 특징을 기술하는 것으로 '정한' 수치가 실제 대상의 의미와 차이를 가지게 마련이라는 것인데, 실제 인식 시스템의 설계 과정에서는 얻을 수 있는 정보량의 한계 및 계산 편의 등을 이유로 하여 이와 같은

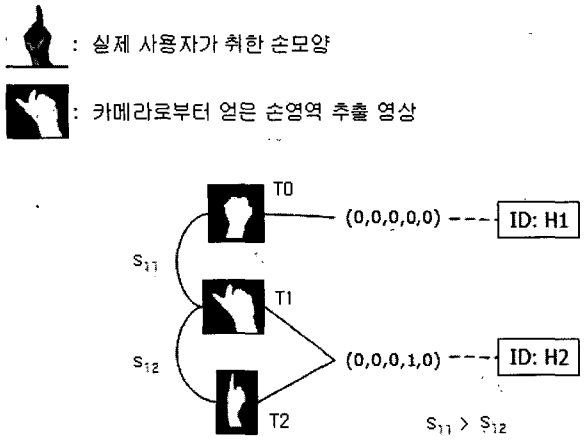

그림 7. 근시안적 탐색의 오류 예
Fig. 7. Example of an error from myoptimal searching.

가정은 일반적이며, 또한 필연적인 것이다. 또 이에 더하여 기본적으로 조명, 배경, 대상 인접 물체 등이 바뀌면 영상에서 추출한 특징 값은 변하게 된다. 즉, 영상처리 과정에서의 잡음을 포함한 여러 요인들로 인해 추출한 영상 특징 값은 안정적이지 않을 수 있다는 것이다. 이러한 오차는 손 모양 인식에 대한 전체적인 성공률은 낮춘다.

다양한 방향 변화에 대한 영상을 통해 방향 변화에 대하여 영향을 받지 않는 인식시스템을 대상으로 했을 때, '장기적인 인식 성능의 개선'이라는 말은 '전체 표본 데이터 전체에 대한 인식 성능 향상'으로 바꾸어 생각할 수 있다. OSS-Net에서는 해당 손 모양 간의 연관성이 노드 간 연결로 정의되며, 결과적으로 손 모양 인식 과정은 노드 간 이동 경로 생성 문제로 귀착된다. 강화학습은 이산적이고(discrete) 확률적인(stochastic) 특성을 갖는 모델을 최적화하는 중요한 이론적 바탕이다.

OSS-Net에서의 손 모양 인식 과정은 각 특징 노드를 경유하여 최종적인 손 모양 특징 노드에 도달하는 경로 탐색 문제에 빗대어 생각할 수 있다. OSS-Net에서 각 노드는 다른 노드로의 여러 연결을 갖는다. 이 때 여러 연결 경로 중, 장기적인 관점에서 높은 성능을 얻을 것으로 예상되는 경로를 선택해야 한다. 따라서 각 노드 간 연결에 대해 '장기적인 측면에서의 높은 성능'을 고려하여 각 노드 사이의 유사도를 정해줄 필요가 있다.

이와 같이 각 노드 간 유사도가 정해진 네트워크상에서의 탐색을 통한 인식 과정은 결정적 마르코프 결정 과정(deterministic Markov decision process)으로 볼 수 있다. 에이전트는 현재의 상태에서 다음 행동을 선택하기 위해 정책 $\pi : S \rightarrow A$ (S: 비연속 상태 공간, A: 비연속 행동)를 학습하게 된다. 임의의 정책 π에 의해서 획득 축적된 보상(cumulative reward)은 식 (2)에 의해 결정된다.[11]

$$V(s,\pi) = r(s,a_\pi) + \gamma \sum_{s'} V(s',\pi) \tag{2}$$

식 (2)에서 r은 에이전트의 보상 함수이고, a_π는 정책 π에 의해서 결정된 행동이며, s'은 다음 상태를, $\gamma \in [0,1]$는 할인율(discount factor)을 의미한다. 마르코프 결정 과정에서 에이전트의 목적은 모든 상태 s에 대해 $V(s,\pi)$를 최대로 하는 정책 π를 학습하는 것이다. 그러한 정책을 최적 정책(optimal policy)이라 하고 π^*로 나타낸다.[11]

$$\pi^* \equiv \arg\max_{\pi} V(s,\pi), \ \forall s \tag{3}$$

어떤 $s \in S$에 대해 다음의 Bellman식에서와 같이 최적 정책 π^*가 존재함은 이미 증명이 된 것이다.[11]

$$V(s,\pi^*) = \max_{a} \left\{ r(s,a) + \gamma \sum_{s'} V(s',\pi^*) \right\} \tag{4}$$

식 (4)에서 $V(s,\pi^*)$는 상태 s에 대한 최적 값(optimal value)이다. 유사도를 결정하기 위한 과정에서 사람이 일일이 결정해주지 않는 한 에이전트들이 보상 함수(reward function)와 상태 변화 함수(state transition function)에 대한 완벽한 지식을 갖는 것은 어렵다. 또 사람이 일일이 지시를 해 준다 하여도 유사도란 설계자의 주관에 크게 좌우될 수 있는 것이므로, 에이전트는 이러한 함수들을 모르고서 환경과의 교류를 통해 직접적으로 최적 정책을 학습할 수 있는 방법이 필요하게 된다.

강화 학습은 그러한 문제를 풀 수 있는 강력하고 실제적인 방법이며, 환경 모델이 불필요한 이와 같은 비모델 강화 학습의 하나가 Q-학습이다. Q-학습의 방법은 식 (5)과 같다.[11]

$$Q(s,a) = r(s,a) + \gamma \sum_{s'} V(s',\pi^*) \tag{5}$$

식 (4)와 식 (5)에 의해 식 (6)이 유도된다.

$$V(s,\pi^*) = \max_{a} Q(s,a) \tag{6}$$

즉, $Q(s,a)$를 통해 최적 정책을 발견할 수 있다.

주어진 OSS-Net 구조를 이용한 손 모양 인식 문제에서 특징 노드 간 유사도를 Q-학습에 의해 정의하기 위하여, 하나의 특징 노드에서 다른 특징 노드로의 이동에 대해 유클리드 거리가 먼 것이 작은 것보다 선택될 가능성이 크다고 가정한다. 그러나 유클리드 거리에 의해 유사하더라도 그것이 반드시 실제 손 모양에 연관되는 특징 노드라는 보장은 없으므로, 잘못된 이동에 대해서는 벌점(penalty)을 준다. 아래에 OSS-Net 학습에 Q-학습을 적용하기 위한 설정을 요약하였다.

Q-테이블의 초기 값 : 즉시 보상
(immediate reward) 값으로 설정
행동(action): 특징 노드 간 이동
상태(state): 각 특징 노드들 간 유클리드 거리와 에이전트들의 이동에 의해 특징 노드 사이 연결에 할당된 값
즉시 보상 (immediate reward)
잘못된 선택일 경우: -5
옳은 선택일 경우:
$Normaize\{1/Euclidean\,Dist\}\times 10$

VI. 실험 및 결과

제안한 OSS-Net을 세 대의 카메라를 이용하는 손 모양 인식 시스템에 적용하여 그 유용성을 검증하여 보았다. 실험은 크게 세 부분으로 구성된다.

영상 특성에 대한 검증
Q-학습 알고리즘을 통한 노드 간 이동 개선 실험
세 대의 카메라로부터 얻은 영상을 동시에 사용했을 때의 손 관찰 시점 변화에 따른 인식 실험

OSS-Net의 구축을 위해서는 영상 특징과 함께 손의 각도 정보가 필요하다. 이를 위해 데이터 취득 과정에서는 세 대의 카메라로 영상을 취득하는 동시에 데이터 글러브를 통해 해당 손 모양의 각도 정보도 함께 취득하였다.

데이터 취득 시에는 그림 8에 보인 5개의 대상 손 모양과 그림 9의 15개의 손 방향을 조합하여 1회당 75개의 데이터를 얻는다. 학습 과정에서 15개의 방향에서 취득한 손 모양을 사용함으로써, 관측 시점 변화에 따른 손 모양 윤곽선의 변화가 인식 과정에 반영될 수 있도록 하였다. 총 3인에 대해 데이터를 얻었으며, 한 사람 당 5회의 반복을 통해 데이터를 얻었다. 이렇게 취득한 총 3375장(3인×3 카메라×5 손 모양×15 시점×5회)의 영상 중 임의로 추출된 1500 장(한 손 모양 당 300장)의 영상을 OSS-Net 구조 생성에 사용하였고 나머지 중 임의로 선택된 1500장(손 모양 당 300장)에 대하여 인식 실험을 수행하였다.

그림 8. 5개의 인식 대상 손 모양
Fig. 8. The five hand-postures to be recognized.

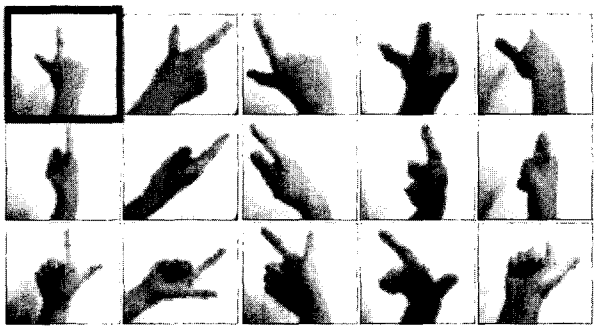

그림 9. 15개의 손 방향
Fig. 9. The fifteen hand-orientations.

그림 10. 손 영역 경계 특징
Fig. 10. Hand-region boundary feature.

1. 영상 특징 실험

2차원 형상 특징으로는 손 영역 경계 특징을 사용하였다. 그림 10에 손 영역 경계 특징을 획득하는 과정에 대해 나타내었다. 취득한 손 모양에서 장축의 방향을 회전시켜 동일하게 맞추고 크기를 정규화한 후 영역의 무게 중심으로부터 경계까지의 거리를 동일한 각도 간격으로 취득한다. 이와 같이 하여 만들어진 n개 항으로 구성된 벡터는 다시 그것을 구성하는 값 중 가장 큰 것을 1로 하여 0과 1사이의 값으로 정규화된다. 실제적인 인식 과정에서 손 영역 추출 영상 간의 비교가 손 영역 경계 특징에 근거하여 행해지므로, 어느 정도로 각도를 세분화하여 기술하여야 이 특성치가 처음 설계 의도대로 손 모양을 잘 구별해줄 것인지에 대한 검증이 필요하다. 이에 대해 실험을 통해 적절한 n을 결정한다.

손 모양 간의 구별 정도를 따질 때에는 "서로 다름"에 대한 가정이 있어야 하는데, 일차적인 입력 영상을 100×100으로 크기 정규화하여 데이터베이스 내 손 모양과 유클리드 거리에 의해 비교한다.

특정 n에 대한 경계 특징의 구분성 정도는 식 (7)과 같이 동일 관찰 시점에 대한 손 모양 변화에 따른 유클리드 거리의 평균 차이와 동일 손 모양에 대한 관찰 시점 변화에 따른 유클리드 거리의 평균 차이의 비로써 구한다.

$$D_n = \frac{\frac{1}{N_h R}\sum_{i\neq j}\sum_{k=1}^{R} \| \underline{x}_{k,i} - \underline{x}_{k,j} \|}{\frac{1}{N_R H}\sum_{p\neq q}\sum_{m=1}^{H} \| \underline{x}_{p,m} - \underline{x}_{q,m} \|} \quad (7)$$

$\underline{x}_{a,b}$: a 방향에서 얻은 b 손모양의 영상으로부터 구한 손 영역 경계 특성 벡터,

$R = {}_{N_r}C_2$, N_r: 시점 변화의 가짓수

$H = {}_{N_h}C_2$, N_h: 고려하는 손 모양의 가짓수

그림 11은 n이 증가함에 따른 손 모양 간의 구별 정도 변화를 그래프로 나타낸 것이다. 너무 작은 n에 대해서는 손 모양 간에 구별이 잘 되지 않고, n이 어느 정도 이상으로 크면 구별 정도가 크게 향상되지 않는 반면 계산에 따르는 소요 시간이 길어진다. 따라서 실험 결과를 통해 100×100 손 영역 영상에 대해 구간의 수 n은 20이 적당한 것으로 결론짓는다.

2. Q-학습 적용 실험

Q-학습을 해당 문제에 적용할 경우, 탐색(Exploration)과 이용(Exploitation) 간의 균형 문제가 제기된다. 본 실험에서 탐색은 유클리드 거리 측면에서 보았을 때 가장 근사하나, 실제로는 서로 다른 손모양일 경우에 대한 고려를 의미한다. 그림 12과 13에 탐색 확률 변화에 따른 평균 인식률과 필요한 반복의 횟수를 나타내었다. 탐색이 이루어 지지 않고 이용만이 이루어지는 경우는 유클리드 거리만을 이용하여 노드 간 이동을 결정하는 방식이다.

탐색 확률이 커짐에 따라 인식률은 개선되고 수렴까지의 속도 또한 빨라짐을 보였다. 그러나 탐색 확률이 지나치게 커질 경우, 오히려 인식률이 급격히 떨어지고 수렴까지의 반복 횟수도 증가한다. 이와 같은 결과에 근거하여 0.55를 탐색 확률로 사용하였다.

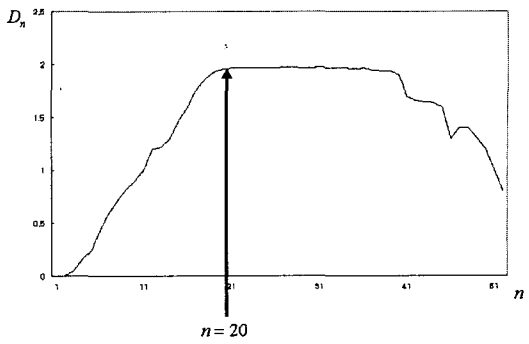

그림 11. 경계 측정 각도 분할 정도에 따른 손 영역 경계 특성의 구분성 정도
(n: 손 모양 경계 특성 추출 시의 각도 분할 수, D_n: n개의 각도로 분할하여 데이터를 취득했을 때의 손 영역 경계 특징 구분성 정도)

Fig. 11. Discriminancy on hand-region boundary feature depending on the sampling number.
(n: the number of sampling, D_n: discriminancy on hand-region boundary feature with n sampling points)

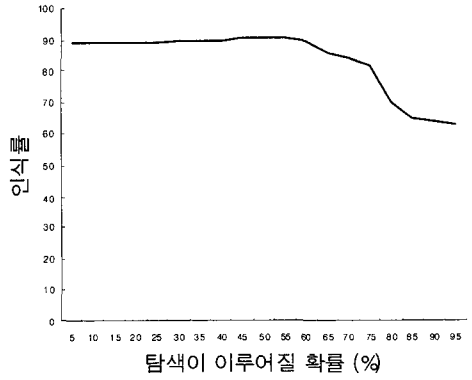

그림 12. 탐색 확률 변화에 의한 인식률 변화

Fig. 12. Recognition rate depending on exploration probability.

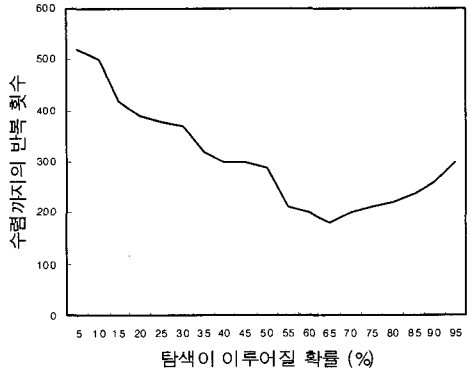

그림 13. 탐색 확률 변화에 따른 수렴까지의 반복 횟수 변화

Fig. 13. Iteration number depending on exploration probability.

3. 세 대의 카메라를 이용한 인식 실험

세 대의 카메라를 이용한 손 모양 인식 시스템에 제안한 방식을 적용하여 보았다. 그림 14에 전체적인 시스템의 흐름도를 보였다. 세 대의 카메라로부터 얻은 세 개의 손 영역 영상으로부터 다시 세 개의 영상 특징 벡터를 추출해내어 손 모양 인식 모듈로 전달한다. 손 모양 인식 모듈은 OSS-Net에서의 탐색과정을 통해 최종적인 인식 결과를 출력한다.

그림 15에 세 대의 카메라로부터 얻은 영상을 이용하여 OSS-Net에서 Q-학습으로 이동 경로를 탐색하는 방식으로 손 모양을 인식했을 때의 실험 결과를 유클리드

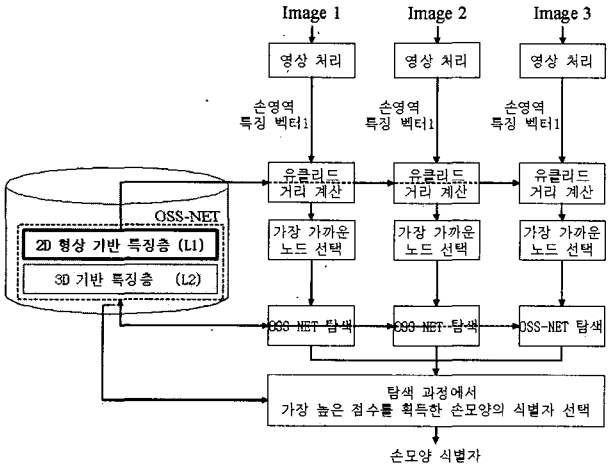

그림 14. 세 대의 카메라를 이용한 손 모양 인식
Fig. 14. Hand-posture recognition using three cameras.

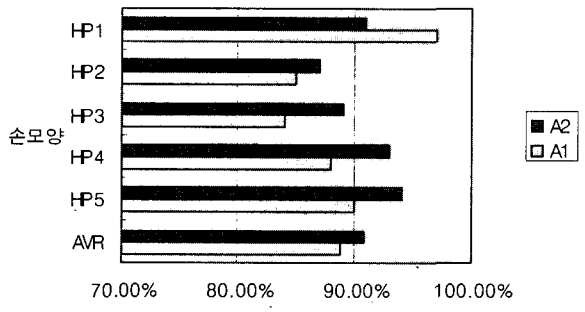

그림 15. 세 대의 카메라를 이용한 손 모양 인식 실험 결과
(A1: 유클리드 거리 유사도에 의해 노드 간 이동을 결정하는 경우의 인식률, A2: Q-학습 결과에 의해 노드 간을 결정하는 경우의 인식률, AVR: 평균 인식률)

Fig. 15. Hand-posture recognition rate using three cameras.
(A1: recognition rate when the transition between nodes is decided by using Euclidean distance, A2: recognition rate when the transition between nodes is decided by Q-learning, AVR: average recognition rate)

거리 비교를 통해 노드 간 이동을 결정했을 때의 경우와 비교하여 나타내었다. Q-학습에 의해 노드 간 이동을 최적화할 경우, 5회의 반복 실험 결과 평균 90.80%의 인식률을 얻을 수 있었다. 각 손모양을 인식하는 데에는 펜티엄4, 3.4GHz 기반의 컴퓨터 환경에서 평균 12.5ms가 소요되었다.

Ⅶ. 결론 및 추후 연구 과제

최근 자연스러운 사용자 인터페이스 수단에 대한 요구가 증대됨과 함께, 손 모양 인식에 대한 연구도 활발히 진행되고 있다. 그러나 손 자체가 많은 관절로 이루어져 형태가 다양하게 변화할 수 있는 물체이고 대개의 경우 방향 변화와 손의 형태 변화가 복합적으로 이루어지기 때문에 손 방향 변화에 대해 불변한 인식 문제는 여전히 어려운 문제로 남아있다. 실제로 여러 손 모양 인식에 대한 연구에서 손 방향은 카메라 정면에서 정해진 각도로 취할 것을 제한 사항으로 두고 있으며, 일상적인 사용자 인터페이스로 사용하는 데에는 한계가 존재한다.

이에 대한 해결 방법으로 본 논문에서는 OSS-Net을 제안하였다. OSS-Net은 2차원 형상 특징과 함께 3차원 구조 특징을 함께 표현하는 구조로 되어 있다. OSS-Net에서의 인식 과정 2차원 형상 기반 특징 층에서의 일차적인 비교와 3차원 구조 기반 특징 층에서 노드 간 이동을 통한 탐색의 과정으로 표현된다. 이러한 구조는 방향 변화에 대한 불변한 인식에 대해 강인한 특징을 갖는다. 우선, 2차원 형상 특징 층에서의 노드들은 하나의 손 모양에 대해 여러 시점에서 관찰한 여러 개의 노드로 존재하여 여러 시점에 대한 고려를 가능하게 하고 또한 3차원 구조 기반 특징 층에서의 노드 간 이동은 동일한 윤곽선 정보를 보이나, 실제로는 다른 손모양일 경우에 대해서도 고려함을 의미한다.

인식의 과정이 노드 간 이동을 통한 탐색 과정에 대응되므로, 한 노드에서 다른 노드로의 이동을 적절하게 정의하는 것이 중요하다. 이를 위해서는 강화학습을 적용하여, 단순히 특성 값이 가장 가까운 것으로 이동하는 것이 아니라, OSS-Net 생성을 위해 주어진 표본 데이터를 바탕으로 전체적으로 가장 높은 성능을 보이는 이동 경로를 얻어, 이것을 노드 간 연결에 적용하도록 하였다. 그 결과, 세 대 카메라 입력 영상을 동시 이용하였을 때 각각 15개 방향에서 취득한 5개의 손 모양을 대상으로 평균 90.80%의 인식률을 얻었다.

오인식은 동일한 윤곽선 형태를 갖는 서로 다른 손 모양이 학습에 사용되었을 경우, 입력된 손 모양이 관측 방향 변화를 통해 동일한 윤곽선을 보일 수 있는 다른 손 모양으로 인식되어 발생하였다. 그러나 학습 과정에서 다양한 방향에서 취득한 손 모양을 사용하는 것은 인식기가 방향 강인성을 갖도록 하기 위하여 반드시 필요하다. 고정된 하나의 방향에서 취득한 손 모양만을 학습에 사용할 경우, 해당 방향에서 취득한 손 모양에 대해서는 높은 인식 성능을 보일 수 있다. 그러나, 본 논문의 실험에서 사용한 동일한 데이터로 실제 테스트해 본 결과, 손 관측 방향 변화에 따른 윤곽선 변화로 인해 50% 이하의 인식률을 보였다. 현재의 성능은 추후 손의 윤곽선 뿐 아니라 경계 정보를 반영하는 영상 특징을 채용함으로써 개선될 수 있을 것으로 예상된다.

이후의 연구 방향은 다음의 두 가지로 요약된다. 첫째, 각 층에서의 노드를 생성함에 있어서 보다 효율적이고 방향 변화에 강인한 특징을 이용하여 성능을 개선시킬 수 있을 것이다. 둘째로, OSS-Net 생성 과정에서 입력된 특성 벡터를 반영하도록 기존 노드를 조정하는 알고리즘의 개선이 필요하다.

참 고 문 헌

[1] Mark Weiser, *The Computer for 21st century*, Sci. Amer., 1991.

[2] Jung-Bae Kim, Kwang-Hyun Park, Won-Chul Bang and Z. Zenn Bien, "Continuous gesture recognition system for Korean sign language based on fuzzy logic and hidden markov model," *Proc. of FUZZ-IEEE*, 2000.

[3] Chan-Su Lee, Sang-Won Ghyme, Chan-Jong Park, and Kwang-Yun Wohn, "Virtual reality software and technology archive," *Proc. of the ACM symposium on virtual reality and technology 1998*, pp.59-65, 1998.

[4] Ernest Gardner, *Gardner-Gray-O'Rahilly anatomy : a regional study of human structure*, Saunders, 1986.

[5] 석동일, *한국 수화의 언어학적 분석*, 박사 학위 논문, 대구대학교, 1989.

[6] J. M. Rehg and T. Kanede, "Visual tracking of high DOF articulated structures: an application to human hand tracking," *Proc. of ECCV'94*, pp.35-46, 1994.

[7] D. Lowe, "Fitting parameterized, three dimensional models to images," *IEEE Trans., PAMI*, vol.13, no.5, pp.441-450, 1991.

[8] B. Moghaddam and A. Penmtland, "Maximum likelihood detection of faces and hands," *Proc. of Int. Workshop on Automatic Face and Gesture Recognition*, pp. 122=128, 1995.

[9] U. Brockl-Fox, "Realtime 3-D Interaction with up to 16 degrees of freedom from monocular video image flows," *Proc. of Int. Workshop on Automatic Face and Gesture Recognition*, pp.172-178, 1995.

[10] H. H. Buelthoff, S. Y. Edelman, and M. J. Tarr, "How are three-dimensional objects represented in the brain?," *A. I. memo no. 1479*, Artificial intelligence lab., Massachusetts Institute of Technology, 1994.

[11] J. Hu and M. P. Wellman, Multiagent reinforcement learning: theoretical framework and an algorithm, Proc. of Int'l Conf. of Machine Learning, 1998.

저 자 소 개

장 효 영(학생회원)
2001년 2월 이화여자대학교 정보통신학과 학사 졸업.
2004년 2월 KAIST 전자전산학과 석사 졸업.
2004년 3월~현재 KAIST 전자전산학과 박사과정.
<주관심분야 : 제스처 인식, 학습이론, 인간-컴퓨터 상호작용>

변 증 남(평생회원)
제 30권 B편 제 10호 참조
현재 KAIST 전자전산학과 교수