인간의 비언어적 행동 특징을 이용한 다중 사용자의 상호작용 의도 분석

Interaction Intent Analysis of Multiple Persons using Nonverbal Behavior Features

윤 상 석, 김 문 상*, 최 문 택, 송 재 복
(Sang-Seok Yun^{1,3}, Munsang Kim¹, Mun-Taek Choi², and Jae-Bok Song³)

¹Center for Intelligent Robotics, Korea Institute of Science and Technology

²Mechanical Engineering, Sungkyunkwan University

³Mechanical Engineering, Korea University

Abstract: According to the cognitive science research, the interaction intent of humans can be estimated through an analysis of the representing behaviors. This paper proposes a novel methodology for reliable intention analysis of humans by applying this approach. To identify the intention, 8 behavioral features are extracted from the 4 characteristics in human-human interaction and we outline a set of core components for nonverbal behavior of humans. These nonverbal behaviors are associated with various recognition modules including multimodal sensors which have each modality with localizing sound source of the speaker in the audition part, recognizing frontal face and facial expression in the vision part, and estimating human trajectories, body pose and leaning, and hand gesture in the spatial part. As a post-processing step, temporal confidential reasoning is utilized to improve the recognition performance and integrated human model is utilized to quantitatively classify the intention from multi-dimensional cues by applying the weight factor. Thus, interactive robots can make informed engagement decision to effectively interact with multiple persons. Experimental results show that the proposed scheme works successfully between human users and a robot in human-robot interaction.

Keywords: human-robot interaction, human intention analysis, multiple-person interactions, confidential reasoning

1. 서론

일상생활 속에서, 다른 사람과 상호작용을 수행하는 인간의 지적 과정을 살펴보면 인간은 오감센서로부터 매 순간 다양한 시공간적 인지 활동을 통하여 상대방의 언어적 비언어적 신호를 습득하고, 추론과 학습 과정을 통하여 상호작용 상황으로부터 대상자에 대한 의도를 파악하고 자의적 의사를 결정하여 그에 따르는 대응 행동을 표현하게 된다[1]. 특히, 사람의 의도 파악은 기능적으로 다른 사람의 행동을 목격함에 의한 뇌 영상 증거를 가지는 심리학적 추론 능력과 정신적 결정에 의해 이루어지게 된다[2].

이러한 사용자의 의도나 감정을 나타내기 위한 다양한 행동 표현 방식에 있어서 언어적 요소에 비해 비언어적 요소는 더욱 효과적이며 높은 전달력을 가지고 있다. 그 이유는 말의 내용보다는 목소리의 강약과 떨림, 시선, 제스처, 억양, 표정, 자세 등이 보다 많은 내면적 정보를 가지고 있기 때문이다[3]. 또한, 사람들 간의 면대면 상호작용에서 상대방의 호감도에 영향을 미치는 요소에 대한 연구에서도 말(7%), 음성

톤(38%), 그리고 몸짓(55%)의 비율로 비언어적 요소가 대부분(93%)을 차지하고 있다[4]. 이러한 비언어적 의사전달은 언어로 표현되는 정보를 보다 정확하게 전달하려는 보조 수단 뿐만 아니라 한정된 언어적 능력 이상의 것을 표현하기 위한 고차원적 의사소통 기능도 수행 가능함을 보여준다[5].

상호작용을 수행하는 로봇이 이러한 복잡한 상황을 이해하고 다양한 지능적 서비스를 제공하기 위해서는 인간의 인지 과정과 유사한 기능 구현이 필요하다. 특히, 많은 사람이 방문하는 박물관이나 쇼핑몰과 같은 동적 환경에서 복수의 사람들을 대상으로 적극적으로 대화 의도를 파악하고 상황에 적절하게 대응하는 로봇은 사용자와의 효과적인 상호작용을 수행하는 데 반드시 필요한 인지적 요소이다. 하지만 사회적 상호작용을 수행하는 대부분의 연구는 교시자에 의한 사회적 학습[6,7]이나 감정 모델에 기반한 사용자와의 정서적 관계[8,9]를 증진하는 데 초점을 두고 있어서 다중 사용자와의 사회적 관계를 형성하는 데는 한계가 존재한다. 따라서, 로봇이 이러한 다수의 사용자를 대상으로 상호작용 의도를 효과적으로 파악하기 위해서는 사용자의 다양한 비언어적인 행동으로부터 신뢰성을 가지는 강인한 인식 및 복잡한 사용자 별 평가 모델이 필수적이다.

로봇의 인식 부분에 있어서, 멀티 모달리티는 단일 모달리티의 상호 보완적 요소를 가지고 있어서 비전과 거리 센서를 이용한 사용자의 위치 검출 및 추적[10], 비전과 음성 센서를 이용한 다중 사용자로부터의 발화자 선정[11] 등 사용자 정보의 신뢰성을 향상시키는데 다양하게 적용되어 왔다. 하지

* 책임저자(Corresponding Author)

Manuscript received March 8, 2013 / revised May 18, 2013 / accepted June 19, 2013

윤상석, 김문상: 한국과학기술연구원 지능로봇 사업단

(yssmecha@gmail.com/munsang@kist.re.kr)

최문택: 성균관대학교 기계공학부(mtchoi@skku.edu)

송재복: 고려대학교 기계공학부(jbsong@korea.ac.kr)

※ 본 연구는 지식경제부 지원으로 수행하는 21C 프론티어 연구개발 사업(인간기능 생활지원 지능로봇 기술개발사업)의 지원으로 연구되었음.

만, 사용자의 의도를 복합적으로 파악하는 부분에 있어서 여전히 한계를 가지고 있다. 최근 3D 깊이 센서의 개발로 인하여 3D 데이터를 이용한 사용자 위치 추정, 공간적 상관관계, 그리고 보다 다양한 사용자의 행동 정보 인식이 가능해졌다.

본 논문에서는 다중 사용자와의 상호작용에서 상호작용 의도를 나타내는 행동에 대한 인식 방법 및 로봇의 응시 결정에 필수적인 사용자 의도 지수를 정량적으로 판단할 수 있는 통합 사용자 평가 모델을 제시한다. 우선, 인간의 상호작용 의도를 나타내는 비언어적 행동에 대한 주요 특성을 분석한 후 비전, 음성, 공간 모달리티를 가지는 센서 정보로부터 다양한 인식 방법을 적용하여 8가지 의도 행동 속성을 추출한다. 그리고, 추출된 개별 인식 정보는 후처리 기법에 의해 신뢰수준을 만족하는 유효한 특징 정보로 통합 사용자 모델에 적용된다. 마지막으로, 다중 추출 정보로부터 개인 별 상호작용 의도를 정량적으로 분류할 수 있는 평가 척도를 제공함으로써 사용자 연산 모델은 사회적 지능 로봇이 다중 사용자와의 상호작용 상황에서 효과적인 관계를 형성하는 데 도움을 주게 된다.

본 논문은 II 장에서는 사용자 행동 특징의 추출 방법을 소개하고, III 장에서는 신뢰추론 기반 사용자 평가 모델을 설명한다. IV 장에서는 실제 환경에서의 실험 결과를 나타내었고, V 장에서 결론을 도출하였다.

II. 사용자 행동 특징 추출

1. 비언어적 행동 특징 분석

상호작용 의도를 나타내는 인간의 행동은 크게 인간-인간 사이의 비언어적 상호작용에서 쉽게 목격될 수 있는 지배적인 기능으로 시각, 동작, 공간, 발성, 촉각의 5가지 카테고리로 구분할 수 있다[12]. 또한, 이러한 요소들은 사용자의 친밀도를 측정하는데 사용되기도 하였다[13]. 본 연구에서는 인간의 4가지 상호작용 행동 카테고리를 바탕으로 로봇 시스템의 인식 능력을 고려하여 8가지 상호작용 의도 속성을 나타내는 주요 행동 특징을 표 1과 같이 구분하였다.

2. 센서 시스템 구성

사용자의 비언어적 행동을 보다 효과적으로 인식하기 위해서는 멀티 센서에 기반한 로봇 인식 시스템의 통합적인 구성이 필요하다. 그림 1에서 보는 바와 같이 카메라 센서는 사람의 눈높이에 맞추어서 4자유도의 목 구동이 가능한 로봇의 머리 중심축에 장착되어 있다. 이러한 구성은 로봇이 인간처럼 면대 면 상호작용을 수행하면서 근거리에서 사용

그림 1. 상호작용 로봇의 센서 시스템 구성.

Fig. 1. System configuration of an interactive robot.

자를 인식하고, 다양한 로봇의 행동 표현이 가능하며, 특정 사용자에게 대한 응시 및 지속적인 사용자 추적 효과를 얻을 수 있다. 다음으로 로봇 베이스를 중심으로 90도 간격의 7.5cm 거리에 위치한 4개의 마이크로폰 조립은 음원 신호를 동시에 측정하고 마이크로폰 사이의 측정 신호차이를 이용하여 전 방향에서의 음원 방향 추정할 수 있다. 지상에서 일정 높이에 고정된 3차원 깊이 센서는 넓은 영역에 대하여 3D 깊이 정보를 획득할 수 있으며 로봇 센서 시스템의 기하학적 구성과 지상 실측 정보(ground truth)를 통하여 로봇을 중심으로 개별 사용자에게 대한 위치 정보 및 신체 정보 추출이 가능하도록 구성하였다.

3. 주요 특징 추출

앞서 구성된 세가지 모달리티를 가지는 멀티 센서로부터 8가지 사용자 행동 정보를 추출하기 위한 과정은 그림 2와 같다. 그리고, 행동 요소 별로 의도를 나타내는 속성에 대한 인식 방법은 다음과 같다.

우선, 시각적 요소는 사용자의 눈의 움직임을 통하여 그 의도를 측정할 수 있다. 하지만 다양한 활동이 이루어지는 공간에서 눈의 움직임을 판별하기란 쉬운 일이 아니다. 대신, 본 논문에서는 Yoon과 Kim에 의해 개발된 정면 얼굴 검출기[14]를 적용하여 사용자의 로봇 응시 여부를 판별한다.

그림 3에서 보는 바와 같이, 얼굴 정면 검출기는 MCT (Modified Census Transform) feature와 Adaboost의 학습 분류에 의해 실시간 사용자의 정면 얼굴 판별이 가능하다. 구성 단

표 1. 행동 요소 기반의 비언어적 특징.

Table 1. Nonverbal features based on behavioral factors.

행동 구분	특징	의도 속성
시각 (Oculesic)	눈의 움직임, x_e	응시하기
동작 (Kinesic)	얼굴 표정, x_f	미소 짓기
	제스처, x_g	손 흔들기
	보행, x_m	다가가기
공간 (Proxemic)	거리, x_p	가까이 서기
	바디 포즈, x_o	마주보기
	상체 기울임, x_l	앞으로 숙이기
발성 (Vocalic)	로봇 호출, x_v	호출하기

그림 2. 멀티모달 센서를 이용한 행동 특징 추출 과정.

Fig. 2. Feature extraction process from multimodal sensors.

그림 3. 정면 얼굴 분류 과정.

Fig. 3. Frontal face classifier.

그림 4. 표정 인식 전처리 과정 및 4가지 얼굴 표정 영상.

Fig. 4. The pre-processed image for facial expression recognition.

계를 살펴보면, 먼저 카메라 입력 영상에서 얼굴을 검출하고 검출된 얼굴 영역 내에서 두 눈을 검출한다. 그리고 검출된 눈 좌표를 기준으로 40x50의 얼굴 영역을 정규화한 후 MCT 변환하여 조명성분을 제거하는 과정을 거치게 된다. 이후 정면과 비 정면 영상으로 학습된 판별 임계 값을 이용하여 정면 얼굴 여부를 판별함으로써 사용자의 시각적 행동 요소에 의한 상호작용 의도를 추정하게 된다.

동작 요소는 사용자의 얼굴 표정, 제스처, 그리고 보행으로 구성된다. 얼굴 표정 인식[15]은 사람의 표정 입력정보를 neural network로 학습된 특징과 비교 매칭하여 4가지 표정 상태(일반, 웃음, 놀람, 화남)로 분류한다. 분류과정을 살펴보면, 초기 입력 영상에 대하여 얼굴 및 눈 검출로 얼굴 영역을 찾은 후 얼굴 영상을 정규화한다. 정규화된 얼굴의 크기는 40x50 픽셀로서 총 2,000 픽셀로 구성된다. 다음으로 정규화된 영상에 대해서 필요 없는 영역을 제거하는 Mask 작업을 수행한 후 각 픽셀의 강도(Intensity)를 신경망의 입력으로 넣어서 4가지의 표정의 특징으로 분류하게 된다. 그림 4는 입력이미지에 대한 전처리 과정 및 최종 분류되는 얼굴의 표정 상태를 나타낸다. 여기에서 사용자의 웃음과 놀람은 상호작용 의도를 표시하는 얼굴 표정 상태이며 화남은 의도를 표시하지 않는 상태로 설정한다.

제스처와 사용자 보행은 OpenNI [16]에서 제공하는 3D 깊이 정보를 이용하여 행동 속성을 추출하였다. 이 라이브러리는 개별 사용자의 위치 정보와 그림 5와 같이 유효한 15개의 스켈레톤 정보를 제공한다. 따라서, 제스처 인식은 다음과 같이 특정 스켈레톤 데이터의 추적으로 구현이 가능하다.

$$\begin{aligned} l_{hand.z} &= P_{hand.z,t} - P_{elbow.z,t}, \\ v_{hand} &= |P_{hand.x,t} - P_{hand.x,t-1}| \end{aligned} \quad (1)$$

그림 5. 15개의 신체 구성 요소.

Fig. 5. The 15 body components.

$$x_g(s) \in \begin{cases} g_{WAV}, & \text{if } l_{hand.z} > 0, v_{hand} > v_{th_h} \\ g_{NWAV}, & \text{otherwise} \end{cases} \quad (2)$$

여기서 $l_{hand.z}$ 는 3D 좌표계 내 Z축에서의 손의 위치 벡터, v_{hand} 는 X축에서의 손의 속도, P 는 각 손과 팔꿈치의 3차원 위치 벡터, g_{WAV} 는 추정되는 손 흔들기 상태, g_{NWAV} 는 손 흔들지 않는 상태, t 는 샘플링 시간, v_{th_h} 는 손의 움직임 속도에 대한 임계 값을 나타낸다. 그리고, 손 흔들기 상태는 의도를 나타내는 행동으로 설정한다.

사용자의 보행 추정은 다음과 같이 사용자의 현재 위치 정보와 이전 위치 정보의 연산을 통하여 이루어진다.

$$\begin{aligned} dx &= P_{torso.x,t} - P_{torso.x,t-1}, \\ dy &= P_{torso.y,t} - P_{torso.y,t-1} \end{aligned} \quad (3)$$

$$v_m = \sqrt{dx^2 + dy^2}, \quad \theta_m = \tan^{-1} \left(\frac{P_{torso.x,t}}{P_{torso.y,t}} \right) \quad (4)$$

$$x_m(s) \in \begin{cases} m_{FW}, & \text{if } v_m \geq v_{th_m}, \theta_m < \theta_{th_m} \\ m_{NFW}, & \text{otherwise} \end{cases} \quad (5)$$

여기서 P_{torso} 는 사람의 현재 위치 벡터, v_m 는 이동 속도, θ_m 는 이동 방향, m_{FW} 는 추정되는 다가가기 상태, m_{NFW} 는 다가가지 않는 상태, 그리고 v_{th_m} 와 θ_{th_m} 는 각 속도와 방향에 대한 임계 값을 나타낸다. 앞서 설정된 보행 의도의 속성에 따라 다가가기가 의도를 나타내는 행동으로 추정된다.

공간 요소는 사용자 포즈, 상체 기울임 각도, 그리고 사용자와 로봇 사이의 공간적 거리에 의해 쉽게 측정할 수 있다. 그림 6은 공간적 요소에 대한 사용자의 특징 추출방법을 나타낸다. 사용자 포즈와 상체 기울임 각도는 다음과 같이 목 P_{neck} , 몸통 P_{torso} , 그리고 왼쪽 어깨 P_{l_sh} 의 3차원 스켈레톤 위치 벡터의 상관관계로 추정할 수 있다.

$$\begin{aligned} l_{human} &= P_{torso} - P_{robot}, \quad l_{body} = P_{neck} - P_{l_sh}, \\ u_1 &= [1 \ 0 \ 0]^T, \quad u_2 = [0 \ -1 \ 0]^T \end{aligned} \quad (6)$$

$$\theta_H = \cos^{-1} \left(\frac{l_{human} \cdot u_1}{\|l_{human}\| \|u_1\|} \right) \quad (7)$$

그림 6. 공간 특징 정보 추출 방법.

Fig. 6. Proxemic feature extraction in 3D coordinates.

$$\theta_B = \cos^{-1} \left(\frac{l_{body} \cdot u_2}{\|l_{body}\| \|u_2\|} \right) \quad (8)$$

$$x_o(s) \in \begin{cases} o_{FR}, & \text{if } |\theta_H - \theta_B| < \theta_{th-o} \\ o_{NFR}, & \text{otherwise} \end{cases} \quad (9)$$

$$l_{lean} = P_{neck} - P_{torso}, \quad u_3 = [0 \ 0 \ 1]^T \quad (10)$$

$$\theta_L = \cos^{-1} \left(\frac{l_{lean} \cdot u_3}{\|l_{lean}\| \|u_3\|} \right) \quad (11)$$

$$x_i(s) \in \begin{cases} l_{LEAN}, & \text{if } |\theta_L| < \theta_{th-l} \\ l_{NLEAN}, & \text{otherwise} \end{cases} \quad (12)$$

여기서 l_{human} 는 사람과 로봇 사이의 위치 벡터, l_{body} 는 사용자의 목과 어깨 사이의 위치 벡터, θ_H 는 X 축을 기준으로 한 사용자의 상대 각도, θ_B 는 사용자의 각도, o_{FR} 는 추정되는 마주보기 상태, o_{NFR} 는 마주보지 않는 상태, l_{lean} 는 목과 몸통 사이의 위치 벡터, l_{LEAN} 는 상체 기울임 상태, l_{NLEAN} 는 상체 기울이지 않은 상태, 그리고 θ_L 는 Z 축을 기준으로 한 로봇의 상체 기울임 각도를 나타낸다. 각 행동 특징에서 마주보기와 상체 기울임 상태가 의도를 나타내는 속성으로 추정된다.

로봇으로부터의 사용자의 거리 특징은 다음과 같이 연산된다.

$$d_p = \sqrt{l_{human,x}^2 + l_{human,y}^2} \quad (13)$$

$$x_p = \frac{n - asc(d_p) + 1}{n} \quad (14)$$

여기서 d_p 는 사람과 로봇 사이의 상대적인 거리, n 는 전체 상호작용 참가자 수, $asc()$ 는 참가자의 거리 오름차순에 대한 순위를 나타낸다.

마지막으로 발생적 요소는 로봇을 기준으로 일정 공간에 위치하는 발화자의 특정 호출에 반응하여 그 음원의 위치를 추정한다[11]. 사용자에게 의해 발화되는 음원은 각 마이크로폰에 이르는 거리에 따라 조금씩 차이가 발생하며, 이 거리 차이가 음원으로부터 각 마이크로폰까지 도달되는 소리들 사이의 시간 지연을 나타내는데 이를 TDOA (Time Delay of Arrival)이라 한다. TDOA는 마이크로폰 어레이의 중심으로부터 음원까지의 거리에 따른 변화는 낮지만 음원의 방위각에 따른 변화는 높다. 여기에서, 추출된 음원 방향은 정확한 발화자를 추정하기 위하여 다음과 같이 검출된 사용자 모델의 위치 정보와 교합(data association) 과정을 거치게 된다.

$$x_v(s) \in \begin{cases} v_{SPK}, & \text{if } s_{TDOA} \geq s_{th}, |\theta_H - \theta_{snd}| \leq \theta_{th-s} \\ v_{NSPK}, & \text{otherwise} \end{cases} \quad (15)$$

여기서 s_{TDOA} 는 TDOA 신호 추정 값, s_{th} 는 TDOA 신호의 임계값, θ_{snd} 는 음원 방위각, v_{SPK} 는 추정되는 발화자에 의한 발화상태, v_{NSPK} 는 조용한 상태, 그리고 θ_{th-s} 는 방위각의 임계값을 나타낸다. 발생적 요소의 의도 추출은 발화자의 발화를 통하여 추정된다.

III. 신뢰 추론 기반 사용자 평가 모델

앞서 추출된 사용자의 행동 특징 정보는 인식기의 특성상 사용자가 실제 상호작용의 의도를 나타내는 행동을 수행하였음에도 불구하고 제대로 인식을 수행하지 못하거나 반대 경우에 의해 오인식 되는 상황이 발생하게 된다.

인식 정보의 통합에 있어서 이와 같은 오인식 정보의 사용을 가급적 배제하고 연속적인 신뢰도 기반의 인식 정보를 적용하기 위하여 개별 인식 이후에 TCR (Temporal Confidence Reasoning) [17]에 기반한 후처리 방법을 적용하였다. 이 방법은 본래 물체 인식에서 연속적인 동일 인식의 결과를 이용하여 신뢰 수준의 수렴 여부를 판단하고 지식체계에 해당 물체의 정보를 등록, 업데이트, 그리고 삭제하는데 사용되어 왔으며 구성은 다음과 같다. 우선, 초기 인식기 A 의 통계적인 평균 인식 성능을 x_A 라고 가정하면, $(1 - x_A)$ 는 오 인식률을 의미한다. 이로부터 $(1 - x_A)^{\tau_A}$ 는 연속적인 인식 상황에서의 인식 결과 τ_A 가 모두 오인식 했을 때 확률로 정의하기 위해서 계산될 수 있는 값이다. 여기서, τ 는 동일 인식 결과에 대한 연속인식의 횟수를 나타낸다. 따라서, 오인식의 확률 $(1 - x_A)^{\tau_A}$ 의 결과 값이 5%내로 수렴하면 그 인식기의 신뢰도 수준이 95%이상을 나타내는 것으로 추론할 수 있다. 예를 들면, 인식기 A 의 인식률이 80%를 나타내면 오 인식률은 20%(0.2)가 된다. 여기에서, 2회 연속 오인식을 나타낼 확률은 4%(0.04)이 내가 되며 이는 95%의 인식 신뢰도에 수렴함을 의미한다. 따라서 최종 인식은 연속적인 2번의 동일 인식으로부터 신뢰수준을 만족하는 결과를 도출하게 된다. 신뢰성을 가지는 데이터의 연속 인식 수는 인식기의 성능에 따라 다음과 같이 연산된다.

$$\tau_A = \min \{ \tau \in I \mid (1 - x_A)^\tau \leq P \} \quad (16)$$

여기서 $P(=0.05)$ 는 신뢰도 수준을 나타낸다.

사용자의 상호작용 의도에 대한 정량적인 평가 모델은 다음과 같다. 후처리 과정을 통하여 취합된 신뢰수준의 사용자 행동 특징 정보는 설정된 의도 속성에 따라 정규화된 값으로 전환된다. 또한, 개별 인식기에 대하여 빈도나 중요도에 따라 다른 가중치를 적용함으로써 보다 효과적인 사용자의 의도 평가가 가능해진다.

$$s = \frac{\sum_{i=1}^n \sum_{t=1}^T \omega_i x_{i,t}(s)}{T} \quad (17)$$

여기서 s 는 사용자의 정량적인 의도 인식 값, n 는 개별 인식기의 개수, ω 는 개별 인식기의 가중치, 그리고 T 는 인식시간을 나타낸다.

복수의 사용자와의 상호작용에서 인간-로봇의 상호작용성을 증가시키기 위한 한가지 방안으로, 로봇은 정량적 평가 모델에 의해 다음과 같이 최고의 의도 인식 값을 가지는 사용자를 선정할 수 있다.

$$s_T = \max[pri(s_1), pri(s_2), \dots, pri(s_n)] \quad (18)$$

여기서 s_T 는 상호작용 참가자 중 최대 의도 표시 대상자, $pri(\cdot)$ 는 개인 별 사용자 의도 평가 값, 그리고 n 은 상호작용에 참가자 수를 나타낸다.

IV. 실험결과 및 고찰

본 장에서는 비언어적 사용자 행동에 대한 인식기의 개별 성능을 테스트하기 위하여 총 10명의 2-30대 성인을 대상으로 로봇으로부터 1.5m의 거리에 위치하고 7가지 행동 특성에 대하여 의도를 나타내는 행동과 나타내지 않는 행동을 각각 10회씩 반복해서 수행하도록 하였다.

그림 7은 각 행동 특징에 대한 개별 인식기의 인식 성능 결과를 보여주고 있다. 정면 얼굴 판별은 총 34,810장의 정면 얼굴과 33,154장의 비 정면 얼굴 이미지로 학습된 특징정보가 사용하였으며 97.4%의 성공률로 최고의 인식 성능을 나타내었다. 표정 인식의 경우에는 총 16,000 이미지(남녀 각 100명, 16개의 조명 변화, 5개의 사용자 얼굴 포즈, 4개의 얼굴 표정)로 학습된 특징정보에 의해 76.2%의 인식 성공률을 나타내었다. 제스처와 보행 인식은 스켈레톤 데이터 추적에 의해 특정 행동 조건의 충족여부로 인식되며 88.2%와 83.1%의 성공률을 나타내었다. 사용자 포즈와 신체 기울임을 통한 의도를 가진 행동 판별은 360개의 사용자 정면 포즈 및 신체 기울임과 360개의 비 정면 바로 선 포즈로부터 설정된 특징 정보에 의해 96.7%와 83.1%의 인식 성공률을 나타내었다.

표 2는 다양한 동적 특성을 가지는 3가지 인식기에 대하여 후처리 기법을 통하여 단일 인식기의 개선된 인식 성능의 결과를 보여준다.

사용자 제스처와 보행의 경우 $\tau=2$ 의 값을 가지며 각각 3.1%와 2.5%의 인식 성능 개선 효과를 나타내었다. 특히, 표정 인식의 경우($\tau=3$)이며 인식 성능이 80.8%로써 신뢰 수준을 만족하는 결과 정보로 인하여 6.1%의 인식 성능 개선 효과

그림 7. 비언어적 행동 특징에 대한 개별 인식 결과.

Fig. 7. Recognition results for nonverbal behaviors.

표 2. 단일 인식기에 대한 TCR 후처리 비교 결과.

Table 2. Comparison results of TCR process for single recognizer.

특징	단일 인식	TCR 후처리	개선률
얼굴 표정	76.2%	80.8%	6.1%
제스처	88.2%	90.9%	3.1%
보행	85.6%	87.7%	2.5%

그림 8. 음원 및 사용자 위치 추정 결과.

Fig. 8. Sound source localization with human model.

과를 나타내었다. 이와 같이 연속적인 인식 결과에 기반한 후처리 과정은 개별 인식기의 오인식을 제거하고 신뢰수준을 만족하는 인식 정보 추출을 통하여 보다 강인한 인식을 가능하게 한다.

로봇 호출 음원에 대한 인식은 평균 55dB의 소음 수준을 가지는 환경에서 로봇으로부터 1.5m의 거리에 위치한 사용자가 일정 각도로 이동한 후 “메로야”라는 단어를 소음 대비 15dB 이상의 강도로 호출하도록 하였다. 그림 8은 사용자 위치에 따른 음원 인식 정보 및 오차를 나타낸다.

사용자 위치의 일정 방위각 이내의 음원 검지율은 73.1%이며 음원의 위치 추정 오차 평균은 0.75° 이며 최대 편차는 9.58° 를 나타내었다. 나머지 26.9%의 음원은 반향음에 의하여 사용자가 위치할 수 없는 위치의 추정결과를 나타내었다. 이에, 로봇은 생성된 사용자 모델에서 10도 이내에 위치하는 실제 발화자를 신뢰성 있게 추적할 수 있게 된다.

그림 9는 복수의 사용자를 대상으로 제시된 사용자 평가 모델의 성능을 평가하기 위해 구성된 행동 시나리오를 나타낸다. 설정된 시나리오는 누구나 쉽게 납득할만한 상호작용 대상자 선정(최대 의도 표시 대상자)을 위해 이벤트 타임라인에 맞추어서 세 명의 사용자(P1~P3)가 일정 시간 간격을 두고 로봇에 접근하고, 개인 별로 설정된 의도를 가진 행동(시각, 동작, 공간, 발성)을 자율적으로 수행한 후 순차적으로 퇴장을 하도록 구성되었다. 또한, 로봇과의 상대 거리 특징에 대한 성능을 점검하기 위하여 타임라인 4에서는 모든 사용자가 일정 시간 동안 상호작용 의도를 나타내지 않는 행동을 수행하도록 하였다. 성능 평가는 각 이벤트 상황에서 로봇의 상호작용 대상자에 대한 응시여부로 판별한다.

총 15명의 사용자가 3명씩 5개조로 구성되어 상호작용 실험에 참가하였다. 그림 10은 이벤트 타임라인에 따라 사용자

사용자	행동	행동 이벤트 타임라인								
		1	2	3	4	5	6	7	8	9
P1	근접 순위 (R)	1	2	3	3	1	1	1	1	1
	x_e									
	x_f, x_g, x_m									
	x_p, x_o, x_i									
	x_v									
P2	근접 순위 (R)		1	1	1	2	3	3		
	x_e									
	x_f, x_g, x_m									
	x_p, x_o, x_i									
	x_v									
P3	근접 순위 (R)			2	2	3	2	3	2	
	x_e									
	x_f, x_g, x_m									
	x_p, x_o, x_i									
	x_v									
최대 의도 표시 대상자		P1	P2	P3	P2	P1	P3	P2	P3	P1

그림 9. 이벤트 순서에 따른 실험 참가자 행동 시나리오.

Fig. 9. Three participants' behavior in event timeline.

그림 10. 비언어적 행동을 이용한 인간-로봇 상호작용 데모.

Fig. 10. Human-robot interaction demo using nonverbal behavior.

그림 11. 이벤트 순서에 따른 실험 참가자의 경로 추적.

Fig. 11. Path trajectories for participants with event timeline.

의 비언어적 행동을 표현하는 데모의 한 장면을 나타내고 있으며, 그림 11은 2D 공간에서 각 타임라인에 따른 3명의 사용자의 경로 추적 결과를 나타낸다.

그림 12. 4가지 상호작용 행동 카테고리에 대한 정량적 평가.

Fig. 12. Quantitative analysis for oculestic, kinesic, proxemic, and vocalic features.

그림 13. 최대 상호작용 의도를 나타내는 대상자 선정 결과 및 로봇의 대상자 응시 추적.

Fig. 13. Experimental results for interlocutor selection representing maximum human intention.

그림 12는 4가지 상호작용 행동 카테고리에 기초하여 3명의 참가자가 표현하는 비언어적 행동에 대한 정량적인 평가 결과를 나타낸다. 다른 행동 특징에 비해 눈 움직임과 제스처의 행동 특징이 상대적으로 많은 의도 변화를 나타내었으며, 이는 목 구동을 이용한 로봇 자체의 응시전환 및 비 대화 대상자의 손 흔들기를 통한 의도 표시에 의한 결과이다. 참가자의 바디 포즈 정보는 좌우로 이동하였던 P1의 상황을 제외하고는 모두 지속적으로 대화 의도를 표현하고 있다.

그림 13은 사용자 평가 모델에 의해 3명의 사용자에 대한 8가지 행동 인식의 정량적인 평가 결과로 선정된 사용자 정보를 보여준다. 이 결과는 그림 9에서의 행동 시나리오 내 상호작용 대상자와 일치함을 보이고 있으며, 전체 실험 참가자에 대한 평가 결과는 93.3%의 일치율을 나타내었다.

요약하면, 목 구동부에 의해 사용자 응시가 가능한 헤드타입 로봇 시스템으로부터 강인한 행동 인식 기반 사용자 평가 모델을 통하여 상호작용 대상자에 대한 실시간 추적 및 상황에 적합한 대상자의 전환이 가능함을 보여주고 있다.

V. 결론

본 논문에서는 로봇이 다수의 사용자와의 효과적인 상호작용을 수행에 필요한 사용자 의도 평가 모델을 개발하였다. 먼저 인지과학의 연구 내용을 토대로 상호작용 의도를 나타내는 사용자의 비언어적 행동 특징을 정립하였고 인식 시스템의 역량을 고려하여 의도를 나타내는 행동 속성을 분류하였다. 주요 행동 특징에 대한 인식은 시각, 청각, 공간의 멀티 센서로부터 제공되는 풍부한 입력 정보에 의해 추출되며, 후처리 과정에 적용된 TCR은 이러한 정보로부터 신뢰도 수준을 만족하는 인식결과를 제공함으로써 개별 인식기의 성능 개선에 효과적으로 작용됨을 확인하였다.

사용자 평가 모델은 동시 다발적인 사용자의 의도를 좀 더 정량적으로 평가하기 위하여 인식 시간 동안 특징 정보를 중요도로 구분하여 해당 행동 특징 결과를 취합하였으며, 복수의 사용자를 대상으로 한 사용자 데모를 통해 실제 상호작용에서 개인 별 의도를 효과적으로 분석할 수 있음을 확인하였다. 추후 과제로는 의도 평가 모델을 기반으로 다양한 상황 정보를 취득하는 방법과 상위 차원의 인지 판단 및 효과적인 의사 결정을 통해 상호작용성을 증진시키는 지능 로봇의 인지 기능 구현에 초점을 두고자 한다.

REFERENCES

- [1] R. B. Murray, M. M. Huelskoetter, and D. O'Driscoll, *Nursing Process in Later Maturity*, Prentice Hall, 1980.
- [2] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Rev., Neurosci.*, vol. 2, pp. 561-567, 2001.
- [3] S. Weis, "Theory and measurement of social intelligence as a cognitive performance construct," Ph.D. Dissertation, Otto-von-Guericke Univ., 2008.
- [4] A. Mehrabian, *Nonverbal Communication*. New York: Aldine-Atherton, 1972.
- [5] M. Argyle, "Non-verbal communication in human social interaction," In R. A. Hinde (Ed.), *Non-verbal communication*. London: University Press, 1972.
- [6] C. Breazeal, J. Gray, and M. Berlin, "An embodied cognition approach to mindreading skills for socially intelligent robots," *Int. J. Robot. Res.*, vol. 28, no. 5, pp. 656-680, 2009.
- [7] A. L. Thomaz and M. Cakmak, "Social learning mechanisms for robots," *Proc of Int. Symp. Robot. Res. (ISR)*, pp. 1-14, 2009.
- [8] G. Castellano, I. Leite, A. Paiva, and P.W. McOwan, "Affective teaching: learning more effectively from empathic robots," *Awareness Magazine: Self-Awareness in Auto. Syst., Inter. Robot.*, doi:10.2417/3201112.003948, Jan. 2012.
- [9] Y.-M. Kim and D.-S. Kwon, "A fuzzy intimacy space model to develop human-robot affective relationship," *Proc. of World Automation Congress*, pp. 1-6, 2010.
- [10] N. Bellotto and H. Hu, "Multisensor-Based Human Detection and Tracking for Mobile Service Robots," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 1, pp. 167-181, 2009.
- [11] B.-G. Lee, J. Choi, S. Yoon, M.-T. Choi, M. Kim, and D. Kim, "Audio-visual fusion for sound source localization and improved attention," *Transactions of the KSME A*, vol. 35, no. 7, pp. 737-743, 2011.
- [12] M. T. Palmer and G. A. Barnett, *Progress in Communication Sciences*, vol. 14. G. A. Barnett (Ed.), Norwood, NJ: Ablex, 1998.
- [13] Y.-M. Kim and D.-S. Kwon, "A fuzzy intimacy space model to develop human-robot affective relationship," *Proc. of World Automation Congress*, pp. 1-6, 2010.
- [14] J. Yoon and D. Kim, "Frontal face classifier using adaboost with mct features," *Proc of Int. Conf. Control, Automation, Robotics and Vision*, pp. 2084-2087, 2010.
- [15] J. Shin and D. Kim, "Expression recognition system on mobile terminals," *Workshop on Image Processing and Image Understanding (in Korean)*, Feb. 2009.
- [16] OpenNI Framework, <http://www.openni.org>
- [17] G. H. Lim and I. H. Suh, "Robust robot knowledge instantiation for intelligent service robots," *Intel. Serv. Robot.*, vol. 3, pp. 115-123, 2010.

윤 상 석

2002년 인제대학교 기계공학부(공학사). 2005년 광주과학기술원 기전공학과(공학석사). 2013년 고려대학교 기계공학과(공학박사). 2005년~현재 한국과학기술연구원 지능로봇 사업단 연구원. 관심분야는 다중 센서 융합, 인간-로봇 인터랙션.

김 문 상

1980년 서울대학교 기계설계학과(공학사). 1982년 서울대학교 기계설계학과(공학석사). 1987년 베를린공대 기계공학과(공학박사). 1987년~현재 한국과학기술연구원 지능로봇 사업단장. 관심분야는 모바일 매니플레이션, 지능로봇 시스템의 설계 및 통합.

최 문 택

1994년 University of Southern California 항공기계공학과(공학사). 1997년 University of Southern California 항공기계공학과(공학석사). 2000년 University of Southern California 항공기계공학과(공학박사). 2003년~2012년 한국과학기술연구원 지능로봇 사업단 책임연구원. 2012년~현재 성균관대학교 기계공학부 조교수. 관심분야는 로봇 지능, 지능로봇 소프트웨어 설계 및 통합.

송 재 복

1983년 서울대학교 기계공학과(공학사). 1986년 서울대학교 기계설계학과(공학석사). 1992년 MIT 기계공학과(공학박사). 1993년~현재 고려대학교 기계공학부 교수. 관심분야는 지능로봇 시스템의 설계 및 제어, 안전 매니플레이터, 이동로봇의 주행.