Journal of Institute of Control, Robotics and Systems (2014) 20(6):612-618
http://dx.doi.org/10.5302/J.ICROS.2014.14.9034 ISSN:1976-5622 eISSN:2233-4335

상호작용 영상 주석 기반 사용자 참여도 및 의도 인식

Recognizing User Engagement and Intentions based on the Annotations of an Interaction Video

장 민 수*, 박 천 수, 이 대 하, 김 재 홍, 조 영 조
(Minsu Jang[1,*], Cheonshu Park[1], Dae-Ha Lee[1], Jaehong Kim[1], and Young-Jo Cho[1])
[1]Electronics and Telecommunications Research Institute

Abstract: A pattern classifier-based approach for recognizing internal states of human participants in interactions is presented along with its experimental results. The approach includes a step for collecting video recordings of human-human interactions or human-robot interactions and subsequently analyzing the videos based on human coded annotations. The annotation includes social signals directly observed in the video recordings and the internal states of human participants indirectly inferred from those observed social signals. Then, a pattern classifier is trained using the annotation data, and tested. In our experiments on human-robot interaction, 7 video recordings were collected and annotated with 20 social signals and 7 internal states. Several experiments were performed to obtain an 84.83% recall rate for interaction engagement, 93% for concentration intention, and 81% for task comprehension level using a C4.5 based decision tree classifier.

Keywords: social interaction, human and robot interaction, social signal, engagement recognition

I. 서론

사람은 타인과 대화를 나누거나 협력할 때 다양한 신호를 통해 정보를 교환하면서 상대방의 상태와 의도를 파악하여 말과 행동에 반영함으로써 자연스러운 상호작용을 이끌어 간다. 상대방이 의기소침한 모습을 보이면 이유를 물어보고 용기를 북돋아주기 위한 말을 해 주며, 상대방이 기쁜 상태이면 함께 기뻐해 준다. 상대방이 인사를 하며 시선을 마주치면 대화를 나누고 싶다는 의사 표시일 수 있다. 만약 인사를 건넸는데 시선을 피한다면 대화하기 싫거나 내가 싫다는 표시일 수 있다. 이와 같이 사람은 항상 상대방의 표정, 행동, 말을 관찰하여 해석함으로써 상대방의 내적 상태를 유추하고 그 내용을 행위에 반영한다.

표정, 행동, 말을 기반으로 한 표현들을 통틀어 소셜신호(Social Signal)라고 한다. 소셜신호는 사람이 사회적 의미 또는 의도를 전달하기 위해 사용하는 관찰 가능한 표현이다[1]. 소셜신호는 일반적으로 언어적 신호와 비언어적 신호로 나뉜다. 언어적 신호는 발화 문장을 통해 직접 정보를 전달하거나 목소리의 특성, 운율, 음향적 단서 등을 통해 간접적으로 전달된다. 비언어적 신호는 다양한 모달리티를 통해 표현되는데 제스처, 자세, 얼굴 표정, 시선 행위 등이 포함된다. 심지어 외형적 특성과 거리두기 행위 등을 통해서도 소셜신호를 전달할 수 있다. 인간 간 상호작용에서 비언어적 신호는 감정, 성격, 규범, 친밀감 등 폭넓은 사회적 정보를 표현한다. 따라서, 비언어적 신호의 의미를 적합하게 이해하는 것은 성공적인 사회적 상호작용을 위해 매우 중요하다[9].

따라서, 로봇이 인간과 자연스럽게 상호작용하려면 로봇도 인간의 표정, 행동, 말을 관찰하고 그 의미를 해석할 수 있어야 한다.

본 연구는 사람의 소셜신호(Social Signal)로부터 내적 상태를 유추하는 기술 개발을 목적으로 상호작용 영상 분석을 통해 수집한 패턴 데이터를 활용한 내적 상태 인식기 개발 내용 및 그 실험 결과를 보이고 그 의미를 논하고자 한다. II 장에서 전반적인 연구 방법을 설명하고, III 장에서 전술한 연구 방법을 기반으로 한 인간-인간 상호작용 실험의 내용과 결과를, IV 장에서 인간-로봇 상호작용 실험의 내용과 결과를 제시한다. V 장에서는 관련 연구와 본 연구의 의미를 논하고, VI 장에서 결론 짓는다.

II. 상호작용 연구 방법

본 연구의 목적은 사회적 상호작용(Social Interaction) 상황에서 사람이 표출하는 다양한 소셜신호(Social Signal)로부터 사람의 내적 상태를 자동으로 유추할 수 있는 인식 모델을 개발하는 것이다. 이러한 목적 달성을 위한 연구 절차를 그림 1에 보인다.

1. 영상 데이터베이스 수집

먼저 상호작용 장면을 촬영하여 영상 데이터베이스를 구축한다. 본 연구에서는 선생님과 학생간 1:1 상호작용을 연구 대상으로 삼았다. 상호작용의 내용은 수학 문제 교습으로 선생님은 순차적으로 수학 문제를 학생에게 제시하고 학생은 문제를 해결한다. 학생이 문제를 해결하는 도중 선생님은 해결 단서를 제공하는 등 도움을 줄 수 있고 학생에게 칭찬이나 격려를 해줄 수 있다. 학생은 선생님에게 도움을 요청하거나 정답을 확인 받을 수 있다. 영상 촬영은 2차에 걸쳐 진행했는데 1차는 사전 연구의 성격으로 사람-사람 간 상호작용 장면을

* Corresponding Author
Manuscript received February 15, 2014 / revised March 15, 2014/ accepted March 30, 2014
장민수, 박천수, 이대하, 김재홍, 조영조: 한국전자통신연구원
(minsu@etri.re.kr/bettle@etri.re.kr/bigsum@etri.re.kr/jhkim504@etri.re.kr/youngjo@etri.re.kr)
※ 본 연구는 산업통상자원부 및 한국산업기술평가관리원의 로봇산업융합핵심기술개발사업의 일환으로 수행하였음[10041659].

Copyright© ICROS 2014

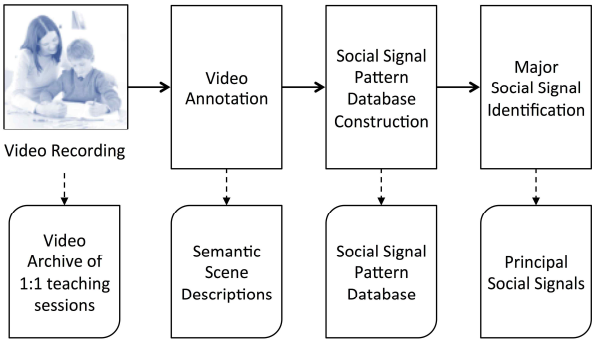

그림 1. 연구 절차[8].
Fig. 1. The process of the study (exerpt from [8]).

촬영하였고, 2차는 인간-로봇 간 상호작용 장면이었다. 로봇이 선생님 역할을 하였고, 사람은 학생 역할을 하였다.

2. 영상 주석(video annotation)

본 연구에서 영상 분석의 목적은 상호작용 참여자의 내적 상태를 인식하기 위한 학습 데이터를 확보하는 과정으로서 영상 내 각 장면에서 관찰되는 소셜신호와 상호작용 참여자의 내적 상태를 주석으로 기록하는 작업을 수행한다. 영상 분석은 인간 관찰자가 수동으로 수행하는데, 일반적으로 영상 주석 도구(Video Annotation Tool)를 활용한다. 본 연구에서는 무료 영상 주석 도구인 ANVIL [2]을 활용하였다. 표 1에 영상 주석의 구조와 의미를 보인다.

3. 패턴 데이터베이스 구축

영상 주석을 기반으로 내적 상태 인식기의 학습과 시험에 활용할 패턴 데이터베이스를 구축한다. 각 패턴은 특정 시간에 관찰된 특정 참가자의 소셜신호와 참가자의 내적 상태를 포함한다. 그림 2는 영상 주석으로부터 패턴 데이터베이스를 추출하여 구축하는 과정이다.

ANVIL을 활용하는 경우 영상 주석 결과는 XML 문서 형식으로 저장된다. XML 문서 안에 표 1에 보인 바와 같은 영상 주석 항목들이 포함되어 있다. XML 파서를 통해 영상 주석

표 1. 영상 주석의 종류와 구조.
Table 1. Video Annotations.

종류	영상 주석	
	구조	의미
소셜신호	*ss(a,T1,T2)*	시간 T1과 T2 사이에 참가자 a의 소셜신호(Social Signal) ss가 관찰되었다.
내적 상태	*is(a,T1,T2)*	시간 T1과 T2 사이에 참가자 a의 내적 상태(Internal State) is가 유추되었다.

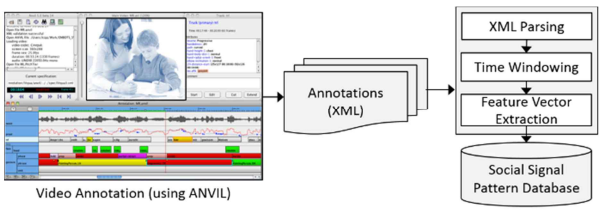

그림 2. 패턴 데이터베이스 구축 절차.
Fig. 2. The process of building a social signal pattern database.

문서를 해석하고 사전에 정한 길이의 시간 구간을 기준으로 해당 시간 구간 내에서 관찰된 소셜신호와 내적 상태값을 기반으로 특징 벡터를 생성하고 특징 벡터들을 모아서 패턴 데이터베이스를 구축한다. 정리해 보면, 패턴 데이터베이스를 구성하는 각 특징 벡터는 특정 시간 구간 내에서 관찰된 소셜신호들에 대한 정보를 포함하며, 클래스 레이블(Class Label)은 내적 상태이다.

특징 벡터 추출: 소셜신호와 내적 상태의 집합을 $S = \{S_1, S_2, S_3, \ldots, S_{N_S}\}$라고 하자. 영상 주석의 결과는 N_S 차원 벡터의 순열 A로 표현할 수 있다. 즉, $A = \{A_t^i | 1 \le i \le N_S \wedge t > 0\}$으로 쓸 수 있으며, A_t^i의 값은 시간 t에서 소셜신호 또는 내적 상태 S_i가 관찰되면 1이고 그렇지 않으면 0이다. 여기서, 시간 t는 연속 시간 상의 한 시점을 가리키나, 일반적으로 영상의 초당 프레임 개수에 의존적이다. 예를 들어, 25fps인 영상이라면, t의 최소 단위는 0.04초가 된다. A로부터 특징 벡터 F를 추출하는 방법은 다음과 같다. A 위에 길이 L인 시간창(Time Window)을 덧씌우고 주어진 시간창 내에서 특정 소셜신호 또는 내적 상태가 관찰되는지 여부를 F에 반영한다. 즉, 특정 시간 t에서 추출한 특징 벡터 F_t는 N_S 차원의 이진 벡터이다. 소셜 신호 S_i에 대하여, F_t^i의 값은 $\sum_{k=t-L}^{t} A_k^i$의 값이 1보다 크거나 같으면 1이고, 그렇지 않으면 0이다. 내적 상태 S_j에 대하여 F_t^j의 값은 $\sum_{k=t-L}^{t} A_k^j$의 값이 $\frac{L}{2}$보다 크거나 같으면 1이고, 그렇지 않으면 0이다. 즉, 내적 상태는 주어진 시간창을 구성하는 영상 프레임의 절반 이상에서 관찰되어야 그 상태가 발생한 것으로 간주한다.

4. 인식기 개발 및 특징 분석

패턴 데이터베이스가 구축되면 내적 상태 인식기를 생성한다. 패턴 데이터베이스는 레이블을 보유한 패턴 벡터들을 포함하므로 지도 학습(Supervised Learning) 모델을 활용하여 내적 상태 인식기를 구축할 수 있다. 본 연구에서는 다양한 패턴분류기 모델들을 활용하여 실험하였으며 그 결과는 III, IV 장에서 설명한다.

III. 인간-인간 상호작용 실험

1. 시나리오와 실험 조건

상호작용 시나리오는 구구단 퀴즈이다. 선생님과 학생이 1:1로 상호작용하며 선생님은 사전에 준비한 문제를 순서대로 제시하고 학생은 문제를 푼다. 선생님은 문제가 인쇄된 문제지를 학생에게 보여주고 학생은 문제지 또는 연습장을 이용하여 문제를 풀 수 있다. 문제 풀이 과정 중 선생님은 힌트를 제공하거나 칭찬, 격려해줄 수 있다. 학생이 문제를 잘 해결하면 선생님은 좀 더 어려운 문제를 선택할 수 있다. 반대로 학생이 문제를 해결하지 못하면 도움을 제공하고 좀 더 쉬운 문제나 관련 문제를 선택하여 제시할 수 있다.

실험에는 3명의 선생님과 6명의 학생이 참여하였다. 선생님 3명 중 2명은 여자 선생님, 1명은 남자 선생님이다. 학생은 남녀 각 3명씩이고, 연령 범위는 8~9세이다.

실험은 학생 당 1번씩 총 6번 수행하였다. 각 실험은 15 ~ 20분 동안 지속되어 총 80여분 길이의 영상을 수집하였다.

2. 영상 주석

영상 주석은 총 2명의 코더(Coder)가 6개의 영상을 나누어

표 2. 영상 주석 대상 소셜신호와 내적 상태.
Table 2. Social Signals and Internal States for Video Annotation.

분류	종류	소셜신호 또는 내적 상태
소셜신호	선생님 발화	질문 대답 반응 기타
	학생 발화	질문 대답 반응 기타
	시선 방향	선생님 문제지 기타
	행동	고개 끄덕이기 고개 흔들기 필기하기 머리 만지기 얼굴 만지기 기타
	자세	꼿꼿하게 앉기 책상으로 상체 기울이기 기타
내적 상태	-	상호작용 참여도(Engagement) 확인 의도(Confirmation Intention)

서로 겹치지 않게 작성하였다. 영상 주석의 내용은 표 2에 보인 바와 같다.

상호작용 참여도는 상호작용 참여자들이 상호작용에 지속적으로 참여하고자 하는 욕구 또는 의지의 수준을 나타낸다. Sidner 등은 참여도를 상호작용 참여자들이 상호 연결성을 관리하기 위해 암묵적으로 사용하는 규약(Protocol)으로 해석하였다[3]. Conte와 Poggi는 참여도를 상호작용 참여자들이 함께 상호작용을 지속하고자 하는 목적에 두는 가치의 수준이라고 정의하였다[4,5]. 특히, Bierman은 사람과 사람 간 상호작용에 존재하는 암묵적으로 합의된 규약을 로봇도 따라주어야만 사람과 자연스럽게 상호작용할 수 있다고 주장하였다[6]. 따라서 상호작용 참여자의 참여도를 정확하게 파악할 수 있다면 보다 자연스럽고 효과적인 상호작용을 이루기 위한 다양한 행위를 계획하여 실행할 수 있을 것이다. 즉, 상호작용 참여도는 가장 중요한 내적 상태 중 하나라고 할 수 있다.

확인 의도는 이 시나리오에 특화된 참여자 상태로서 촬영된 영상을 분석하면서 발견하였다. 학생들은 선생님이 제시한 문제를 해결하거나 선생님이 힌트를 제공하는 과정에서 한 질문에 대한 대답을 할 때 짧게 선생님 쪽을 바라보는 행동을 자주 보여주었다. 이러한 행동은 선생님과 영상 코더들에게 '문제 풀이 또는 대답이 자신 없을 때 선생님에게 확인의 신호를 받고자 하는 의도가 담긴 행위'로 인식되었고, 이를 확인 의도(Confirmation Intention)로 명명하였다. 만약 로봇이 학생의 이러한 행동 뒤에 숨겨진 확인 의도를 유추해 낼 수 있다면 학생이 필요로 하는 응답이나 신호를 제공함으로써 상호작용을 자연스럽게 이끌어 나갈 수 있을 것이다. 본 연구에서 내적 상태는 모두 이진값으로 기록하였다. 즉, 참여도의 경우 '참여함(1)'과 '참여안함(0)'으로 기록하였고, 의도의 경우 '의도 관찰됨(1)'과 '의도 관찰 안됨(0)'으로 기록하였다.

소셜신호는 사람의 얼굴 표정, 행위, 자세, 발화 내용으로부터 직접 관찰 가능하므로 그 내용을 직접 기록하면 된다. 하지만, 내적 상태는 직접 관찰할 수 없으므로 영상 코더의 주관적인 판단이 필요하다. 즉, 내적 상태는 코더에 따라 상황을 해석하는 기준이 서로 다르다는 전제가 있으므로 객관적인 데이터를 확보하기 어려울 수 있다.

3. 실험 결과

II 장에서 설명한 패턴 벡터 추출 방법을 이용하여 영상 주석으로부터 패턴 데이터베이스를 구축하였고, 다양한 패턴 분류기 모델을 대상으로 시간창의 길이 L을 1초 ~ 6초까지 변화시키며 상호작용 참여도와 확인 의도에 대한 인식 모델을 구축하고 인식 성능을 측정하였다. 그 결과는 그림 3과 같다. 인식 성능은 10차 교차 검증을 통해 산출하였다.

시간창의 길이에 따른 인식 성능을 살펴 보면, 참여도의 경우 C4.5는 시간창의 길이가 길어짐에 따라 일관되게 인식률이 저하되었고, NB, MLP는 시간창과 인식률 간 상관관계가 높지 않아 보였으며, RF의 경우에는 시간창이 길어짐에 따라 꾸준하게 인식률이 개선됨을 볼 수 있다. 확인 의도 인식 성능의 경우에는 모든 패턴 분류기에 대해 일관되게 시간창이 길어짐에 따라 인식률이 저하되는 현상을 발견하였다. 이러한 현상의 이유를 밝혀내기 위해서는 학습 데이터에 대한 체계적인 분석 작업이 필요하겠으나, 본 고에서는 인식 결과를 기반으로 그 원인을 유추하는 선에서 결과를 해석해 보고자 한다.

확인 의도는 선생님의 얼굴을 바라보기 위해 시선을 이동하는 행위나 자세를 변화시키는 등 순간적으로 관찰되는 소셜신호의 조합에 의해 판단된다. 즉, 사건 중심적이고 단발적인 행동을 기반으로 판단되는 내적 상태이다. 한편 참여도는 선생님이 질문을 하거나 설명을 할 때 선생님을 바라보기 위해 시선을 이동하거나 자세를 고쳐 앉는 등의 사건 중심적 단서들뿐 아니라 선생님이 설명하는 동안 지속적으로 선생님과 눈을 마주친다거나 질문을 하는 등 시간 지속적인 단서들을 근거로 판단되는 내적 상태이다. 참여도는 단서의 지속성 측면에서 복합적인 성격을 띠는 셈이다. 따라서, 확인 의도의

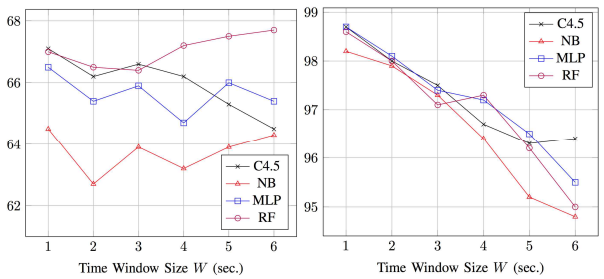

그림 3. 패턴 분류기에 따른 인식 성능 (좌: 참여도, 우: 확인 의도, C4.5: 결정트리 모델, NB: Naïve Bayes 분류기, MLP: 다층퍼셉트론, RF: Random Forest).

Fig. 3. The recall performance for classifiers (left: engagement, right: confirmation intention, C4.5: Decision Tree, NB: Naïve Bayes Classifier, MLP: Multilayer Perceptron, RF: Random Forest).

경우 시간창이 길면 관찰되는 소셜신호 단서들이 더 많이 축적되어 결과적으로 의도 판단을 모호하게 하는 요소들이 늘어나므로 인식 성능이 저하된다고 판단할 수 있다. 반면

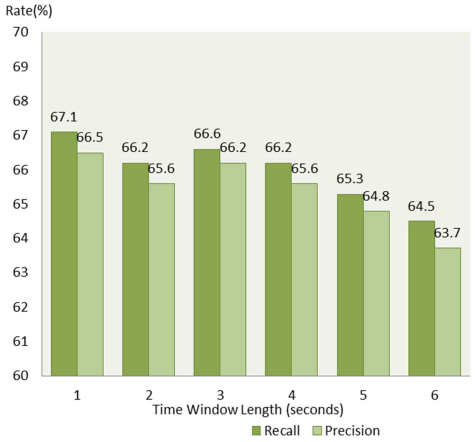

그림 4. 시간창의 길이에 따른 참여도 인식 성능.

Fig. 4. The performance of engagement recognition for varying lengths of time window.

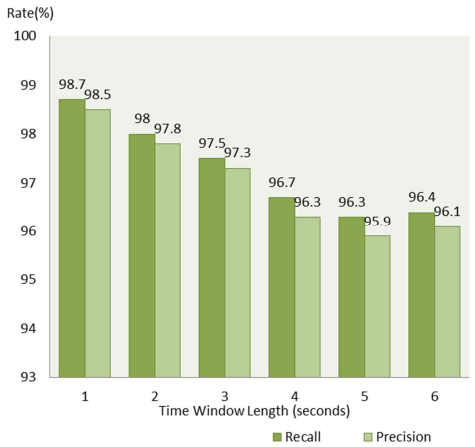

그림 5. 시간창의 길이에 따른 확인 의도 인식 성능.

Fig. 5. The performance of confirmation intention recognition for varying lengths of time window.

표 3. 학습된 결정 트리의 예.

Table 3. A decision tree trained by annotation data.

```
posture.stoop = 0
|   studentSpeech.answer = 0
|   |   teacherSpeech.reaction = 0
|   |   |   teacherSpeech.question = 0
|   |   |   |   posture.straight = 0: 0 (3753.0/1087.0)
|   |   |   |   posture.straight = 1: 1 (9.0)
|   |   |   teacherSpeech.question = 1
|   |   |   |   behavior.hairtouch = 0: 1 (711.0/311.0)
|   |   |   |   behavior.hairtouch = 1: 0 (17.0/4.0)
|   |   teacherSpeech.reaction = 1: 1 (503.0/206.0)
|   studentSpeech.answer = 1
|   |   behavior.hairtouch = 0: 1 (701.0/220.0)
|   |   behavior.hairtouch = 1: 0 (17.0/6.0)
posture.stoop = 1: 1 (121.0)
```

참여도의 경우는 인식 성능과 시간창의 길이 간 상관 관계가 낮다고 볼 수 있다.

패턴 분류기 중에서 참여도와 확인 의도에 대해 모두 상대적으로 안정적이고 높은 성능을 보인 모델은 C4.5 알고리즘 기반의 결정트리 모델이었다. C4.5 모델을 적용한 인식 성능의 상세한 결과는 그림 4와 그림 5에 보인다.

표 3은 참여도 인식을 위한 패턴 분류기 학습의 결과로 생성된 결정 트리이다. 20개의 모든 속성을 포함하는 트리는 크기가 너무 크므로, 여섯 개의 속성만을 기반으로 학습한 트리를 보였다. 트리를 보면 상체가 바른 상태(posture.stoop=0)에서 대답(studentSpeech.answer=1)을 하는 경우 머리를 만지지 않는(behavior.hairtouch=0) 상태이면 참여도가 높다고 결정한다. 선생님이 학생에게 반응(teacherSpeech.reaction=1)하는 상황이거나, 학생이 질문을 하되(studentSpeech.question=1) 머리를 만지지 않는 상황(behavior.hairtouch=0)이면 참여도가 높다고 판단한다. 이와 같이 주석 데이터로부터 학습된 결정 트리는 상식적으로 수긍할 만한 판단 경로를 포함함을 발견할 수 있다. 그러나 최상위 노드를 보면, 상체 자세가 꾸부정한 경우(posture.stoop=1) 참여도가 높다고 결정하도록 되어 있어 데이터 기반 추출의 한계점을 보이기도 한다. 다만, 본 트리는 여섯개의 속성만을 이용해서 생성한 트리이므로 20개 속성 전부를 이용하는 경우 이러한 문제는 보다 줄어들게 된다.

IV. 인간-로봇 상호작용 실험

1. 시나리오와 실험 조건

III 장의 인간-인간 상호작용과 동일한 시나리오를 활용하였다. 단, 인간-로봇 상호작용 실험에서는 로봇이 선생님의 역할을 수행하였다. 인간-로봇 상호작용이므로 보다 정형화된 수학 퀴즈 시나리오를 구성하였다. 퀴즈는 난이도에 따라 총 3단계로 구성되며 각 단계는 5개의 퀴즈를 포함한다. 시나리오는 가장 쉬운 단계의 문제로부터 시작하며 점차 어려운 단계의 문제로 넘어간다.

실험을 위한 시스템 구성은 그림 6과 같다. 퀴즈의 내용은 로봇 우측의 대형 텔레비전과 학생 앞 테이블에 놓인 타블렛에 동시에 제시된다. 로봇 플랫폼은 EVeR-4를 활용하였다[7].

로봇은 언어뿐 아니라 표정, 동작 등 비언어적 행위도 표현하였다. 로봇 시스템 조작, 퀴즈 문제 제시, 학생의 응답 해석은

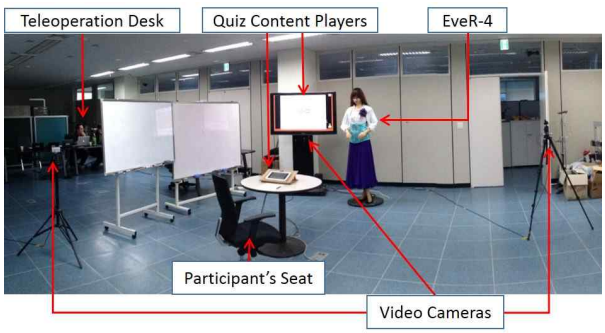

그림 6. 인간-로봇 상호작용 실험 환경과 시스템 구성.

Fig. 6. The environment and system settings for human-robot interaction experiments.

표 4. 주석으로 기록한 내적 상태 목록.
Table 4. List of internal states for annotation.

종류	내적 상태	의미
상태	상호작용 참여도	참여자가 로봇과 상호작용에 참여하는지 여부
	과업 몰입도	참여자가 퀴즈 문제 풀이 과업에 몰입하고 있는지 여부
	이해도	참여자가 퀴즈 문제를 얼마나 수월하게 해결하는지 여부
의도	확인 의도	참여자가 로봇 선생님에게 정답 확인 받고자 하는 의도
	집중 의도	참여자가 과업에 집중하고자 하는 의도
	긍정 의도	참여자가 로봇 선생님의 제안에 긍정하는 의도
	부정 의도	참여자가 로봇 선생님의 제안에 부정하는 의도

모두 원격조정에 의존하였다. 원격조정은 로봇과 학생이 상호작용하는 장소와 차단된 공간에서 이루어졌다.

실험 전 참가자에게 퀴즈 과정과 응답 방법을 설명하였고, 간단한 로봇 시연을 보여주어 낯섦을 완화시킬 수 있도록 하였다.

피실험자는 총 7명으로 여자 5명, 남자 2명이다. 연령대는 10 ~ 11세로서 평균 10.4세, 표준편차는 0.49이다. 각 참가자별로 한번씩 총 7번의 실험이 시행되었다. 각 세션은 8 ~ 12분 길이이며, 총 75분 분량의 영상 데이터가 수집되었다.

2. 영상 주석

본 실험에서는 표 2의 소셜신호와 내적 상태 외에 인간-로봇 상호작용을 위해 필요한 추가적인 내적 상태들을 주석으로 기록하였다. 본 실험에서 주석으로 기록한 모든 내적 상황을 표 4에 보인다.

긍정과 부정 의도는 한 개의 퀴즈 문제 풀이가 종료한 후 로봇이 퀴즈를 더 풀어보겠냐는 제안을 할 때 학생의 반응 행동으로부터 유추되는 의도이다.

영상 주석은 코더 세 명에 의해 기록되었다. 각 코더는 7개의 영상을 모두 분석하였다. 따라서, 총 세 벌의 중복 주석 데이터베이스가 구축되었다. 코더별로 기록한 주석의 일치도를 kappa 지수를 구하여 평가하였다. 평가 결과 자세, 행동, 시선 방향, 로봇 발화에 대해서는 kappa 지수값이 0.6 이상이어서 나쁘지 않은 일치도를 보였다. 그러나, 모든 내적 상태를 포함한 그 외의 소셜신호들에 대해서는 kappa 지수값이 0.6 이하였으며 최하 0.3까지 기록하여 낮은 일치도를 보였다.

3. 실험 결과

III 장에서의 실험 결과를 토대로 본 실험에서는 C4.5 결정트리 모델만을 사용하여 인식 성능 실험을 수행하였다. 시간창의 길이에 따른 상호작용 참여도 인식 성능은 그림 7과 같다.

시간창의 길이가 1초일 때 가장 높은 인식 성능을 보임을 볼 수 있다. 시간창의 길이를 1초로 고정하여 실험한 내적 상태 인식 성능은 그림 8과 같다.

상호작용 참여자의 의도 인식 성능은 그림 9와 같다.

코더에 따른 내적 상태 인식 성능의 차이를 비교해 보기 위해 상호작용 참여도를 대상으로 실험하였다. 그 결과는 그림 10과 같다. 그림 10에서 Coder1은 영상 코더 중 한 명의 아이디이며, Coder1의 인식 성능은 Coder1의 영상 주석 데이터만을 대상으로 10차 교차검증 방식으로 상호작용 참여도 인식 실험을 수행한 결과이다.

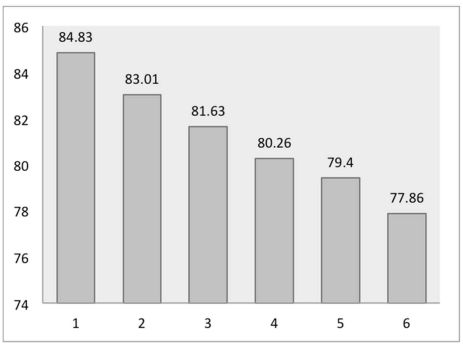

그림 7. 시간창의 길이에 따른 상호작용 참여도 인식률 (시간창 길이: 1초 ~ 6초).
Fig. 7. The recall rates per different time window lengths for interaction engagement (time window length: 1 ~ 6 sec.).

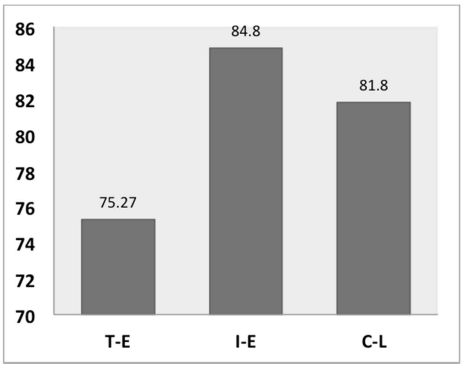

그림 8. 내적 상태 인식 성능 (T-E: 과업 몰입도, I-E: 상호작용 참여도, C-L: 과업 이해도, 시간창 길이 = 1초).
Fig. 8. The recall rates for internal states (T-E: Task Engagement, I-E: Interaction Engagement, C-L: Comprehension Level, Time Window Length = 1 sec.).

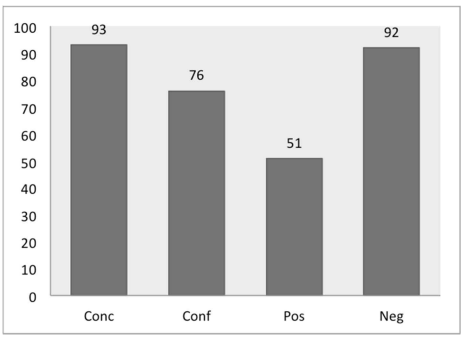

그림 9. 의도 인식 성능 (Conc: 집중 의도, Conf: 확인 의도, Pos: 긍정 의도, Neg: 부정 의도, 시간창 길이 = 1초).
Fig. 9. The recall rates for intentions (Conc: Concentration Intention, Conf: Confirmation Intention, Pos: Positive Intention, Neg: Negative Intention, Time Window Length = 1 sec.).

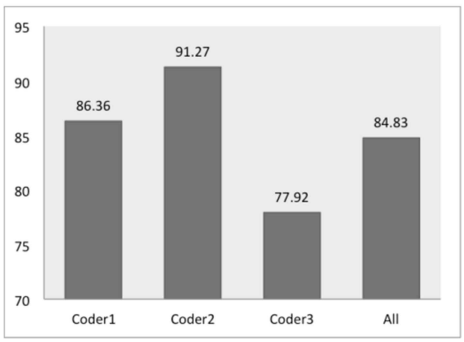

그림 10. 영상 코더에 따른 참여도 인식 성능.
Fig. 10. Coder-dependent performance of engagement recall.

그림에서 보듯이 코더에 따라 인식 성능의 차이가 존재한다. Coder3의 인식 성능이 떨어진다는 것은 그만큼 Coder3의 영상 주석 데이터가 자체적인 일관성이 낮다는 사실을 암시한다.

V. 관련 연구 및 논의

인간-로봇 상호작용 분야의 관련 연구는 로봇의 사회적 행위 표현이 사람의 로봇과 상호작용에 대한 인식을 어떻게 변화시키는지 실험을 통해 규명하는 연구들이 주를 이루어왔다. 자동화된 패턴인식 기술을 기반으로 상호작용 참여자의 내적 상태를 자동으로 인식하는 방법에 대한 연구는 많지 않다.

영국 버밍햄 대학의 Castellano 등은 iCat 로봇과 어린이 간의 상호작용 중 어린이의 참여도 인식 방법을 연구했다[6]. 상호작용 시나리오는 체스 게임이었고, 연구 목적은 어린이의 소셜신호 외에 게임 진행 상황 정보를 패턴 정보로 활용했을 경우 참여도 인식 성능이 어떻게 변화하는지 규명하는 것이었다. 총 8명의 어린이 참가자를 대상으로 16개의 영상을 수집하였으며, 영상들을 7초 길이의 영상 조각으로 분리한 뒤 의미있는 조각 96개를 골라서 인식 실험에 사용하였다. 특징은 소셜신호로는 어린이의 표정(미소 여부), 시선방향, iCat의 감성 반응 여부를 포함하고 맥락 정보로는 게임 상태 정보를 포함하였다. 인식기 모델은 베이지안 네트워크(Bayesian Network)를 활용하였다. 실험 결과 참여도 인식 성능은 소셜신호만 활용했을 경우 93.75%, 맥락 정보만을 활용했을 경우 78.13%, 두 정보를 모두 활용했을 경우 94.79%였다. 이 연구는 높은 인식 성능을 결과로 얻었지만, 패턴 데이터의 개수가 너무 적기 때문에 결과를 일반화하여 받아들이기 어렵다. 베이지안 네트워크와 같은 확률통계 기반의 패턴분류기는 더욱 많은 학습 데이터가 필요하기 때문에 더욱 그러할 뿐 아니라, 7초 길이의 영상 조각들 중 저자들이 선택한 96개만을 사용했다는 점은 문제로 지적할 수 있다. 본 논문은 시간창의 길이가 1초일 때 총 13,460개의 패턴 벡터를 추출하였고 추출된 모든 패턴을 학습에 적용하였으며, 약 84.83%의 참여도 인식 성능을 얻었다.

Castellano 등은 [7]에 이어 사회적 맥락과 과업 맥락이 참여도 인식에 미치는 영향을 거듭 연구했다[10]. 이 연구도 체스 게임 시나리오를 기반으로 어린이를 대상으로 실험했다. 총 40명의 어린이가 참여하여 iCat과 체스 게임을 한 후 설문을 통해 게임 중 참여도가 어느 정도였는지 스스로 기술하도록 하였다. 각 영상으로부터 자동으로 게임 상태, 게임 진행 상태, 어린이의 감정 인식 결과, 게임 승패 결과, 체스 말의 움직임 횟수, 게임 시간을 특징으로 추출하여 인식기를 생성하였다. 인식 모델은 SVM(Support Vector Machine)을 적용하였고, 최고 80% 성능을 기록하였다. 이 연구는 한 게임의 영상 전체를 입력으로 참여도를 예측하는 방식을 취하므로 게임 중 사용자의 참여도 상태를 인식하여 반응하는 상호작용이 불가능하다. 본 논문의 접근 방법은 정해진 길이의 시간창을 이동하면서 시간창 안에서 참여도를 인식하므로 사용자의 참여도에 따라 성공적인 상호작용을 위해 적절한 반응을 제공할 수 있다는 장점이 있다. 인식률에 있어서도 본 논문은 84.83%로 더 높은 성능을 기록하였다.

Kapoor 등은 컴퓨터 기반의 튜터링 시스템을 기반으로 학생의 좌절(Frustration) 상태를 예측하는 방법을 제안하였다[11]. 이 방법은 고개 움직임, 표정, 피부 전도도, 마우스 버튼 압력을 특징으로 사용하였다. 이 방법은 다수의 센서를 장착한 의자와 피부 전도도 측정 장치 등 침습적 장치를 활용하였기 때문에 실용성은 떨어진다.

본 연구는 상호작용 참여도뿐 아니라 과업 몰입도, 이해도 등의 내적 상태와 집중 의도, 확인 의도, 긍정 의도, 부정 의도 등 다양한 행위 의도를 대상으로 실험 결과를 제시함으로써 기존 연구에 비해 내적 상태 인식의 폭을 넓혔다. 인식 성능은 상호작용 참여도 84.83%, 집중의도 93% 등 비교적 우수한 성능을 기록하였으나, 긍정의도 51%, 과업 몰입도 75.2% 등 낮은 인식률을 기록하기도 하였다.

인식 성능의 의미와 활용도를 논하기 이전에 패턴 데이터의 특성을 살펴볼 필요가 있다. 내적 상태와 의도는 상호작용 중 한정된 횟수로 잠깐동안 존재하는 경우가 대부분이다. 따라서 시간창을 이동시키며 연속적으로 추출한 패턴은 분포가 심하게 치우치게 된다. 예를 들어, 상호작용 참여도의 경우 높은 참여도로 표시된 패턴 개수는 10,364개인데 반해 낮은 참여도로 표시된 패턴 개수는 3,096개이다. 의도의 경우는 더욱 심하다. 의도 없음으로 표시된 패턴의 개수는 11,489개인데 반해 집중 의도는 1,055개, 확인 의도는 277개, 긍정 의도는 95개, 부정 의도는 544개이다. 긍정 의도의 인식률이 유독 낮은 이유를 이를 통해 짐작해 볼 수 있다. 따라서, 이러한 심한 불균형이 내재된 데이터를 효과적으로 처리할 수 있는 특징 벡터 추출과 학습 방법을 연구할 필요가 있다.

마지막으로, 영상 주석을 통해 내적 상태를 기록하는 작업은 코더의 주관적 판단과 코딩 작업을 수행하는 상황에 따라 매우 큰 편차를 보이므로 객관성이 적다. 따라서, 높은 객관성을 확보하기 위한 다양한 시도가 필요하다. 예를 들어, Castellano가 시도한 바와 같이 영상 조각을 제작하고 이를 다수의 사람들에게 배포하여 내적 상황을 기록하도록 할 수 있다. 이와 같은 크라우드 소싱(Crowdsourcing) 방법은 중복 주석의 개수를 대량화함으로써 통계적으로 유의한 객관성을 확보하는데 유리할 수 있다.

VI. 결론

본 논문은 상호작용 장면 영상 분석을 통해 소셜신호로부터 상호작용 참여자의 내적 상태를 유추할 수 있는 인식기의 개발 과정과 실험 결과를 보였다. 총 20여개의 소셜신호를 입력으로 상호작용 참여도, 과업 몰입도, 과업 이해도, 집중

의도, 확인 의도, 긍정 의도, 부정 의도를 대상으로 인식 실험을 수행하였으며, 상호작용 참여도는 84.83%, 집중의도는 93%의 우수한 성능으로 인식하였다.

본 논문의 접근 방식은 인간 코더가 영상을 관찰하여 기록한 주석을 기반으로 하므로 실제 운용할 수 있는 인식 시스템을 구축하는데 직접 적용할 수 없다. 화행(Speech Act)을 포함하여 본 논문에서 활용한 대부분의 소셜신호들은 자동인식이 어려운 패턴들이기 때문이다. 따라서, 향후 본 연구를 통해 수집한 영상 데이터를 대상으로 각종 소셜신호 인식기를 직접 적용하고 그 인식 결과를 기반으로 내적 상태를 인식하기 위한 연구를 수행할 계획이다. 또한, 영상 코딩의 객관성 담보를 위해 크라우드소싱(Crowdsourcing)을 기반으로 한 영상 주석을 시도하여 그 결과의 객관성을 검증하는 작업도 수행하고자 한다.

REFERENCES

[1] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli, "Social signal processing: The research agenda," In T. B. Moeslund, A. Hilton, V. Krger, & L. Sigal (Eds.), *Visual Analysis of Humans*, pp.511-538, 2011.

[2] M. Kipp, "Multimedia annotation, querying and analysis in ANVIL," *In: M. Maybury (ed.) Multimedia Information Extraction, Chapter 19*, IEEE Computer Society Press.

[3] C.L. Sidner et al., "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, pp. 140-164, 2005.

[4] R. Conte and C. Castelfranchi. "Cognitive and social action," University College London, 1995.

[5] I. Poggi, "Mind, hands, face and body, A goal and belief view of multimodal communication," Weidler, Berlin, 2007.

[6] G. Castellano et al., "Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features," 2009.

[7] H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, and H. Lee, "Uses of facial expressions of android head system according to gender and age," *In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pp. 2300-2305.

[8] M. Jang, D. H. Lee, J. Kim, and Y. Cho, "Identifying principal social signals in private student-teacher interactions for robot-enhanced education," *In RO-MAN, 2013 IEEE*, pp. 621-626, Aug. 2013.

[9] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 17431759, 2009.

[10] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in HRI: An exploration of social and task-based context," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 421-428, 2012.

[11] A. Kapoor, W. Burleson, and R. W. Picard. "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724736, 2007.

장 민 수

1988년~1992년 서강대학교 전자계산학과(학사). 1992년~1994년 서강대학교 대학원 전자계산학과(석사). 1999년~현재 한국전자통신연구원 책임연구원. 2005년~현재 KAIST 전산학과 박사과정.

박 천 수

1999년~2001년 충남대학교 컴퓨터과학과(석사). 2013년~현재 충남대학교 컴퓨터공학과 박사과정. 2001년~현재 한국전자통신연구원 선임연구원.

이 대 하

1999년~2001년 경북대학교 대학원(석사). 2001년~현재 한국전자통신연구원 선임연구원.

김 재 홍

1990년~1994년 경북대학교 컴퓨터공학(학사). 1994년~1996년 경북대학교 컴퓨터공학(석사). 1996년~2006년 경북대학교 컴퓨터공학(박사). 2001년~현재 한국전자통신연구원 (실장, 책임연구원).

조 영 조

1979년~1983년 서울대학교 제어계측공학과(학사). 1983년~1985년 KAIST 전기및전자공학과(석사). 1987년~1989년 KAIST 전기및전자공학과(박사). 2004년~현재 한국전자통신연구원 (단장, 연구위원, 책임연구원).