행동 및 음성인식 기술을 이용한 대화형 스마트 쿠킹 서비스 시스템 개발

문유경*, 김가연*, 김유하*, 박민지*, 서민혁**, 나정은***
*연세대학교 응용통계학과
**연세대학교 컴퓨터과학과
***연세대학교 학부대학
moonyg08300@gmail.com, gomigom1068@yonsei.ac.kr, yousmileforever@yonsei.ac.kr,
minjipark214@gmail.com, dbd05088@naver.com, jenah@yonsei.ac.kr

Development of an interactive smart cooking service system using behavior and voice recognition

Yu-Gyeong Moon*, Ga-Yeon Kim*, Yoo-Ha Kim*, Min-Ji Park*, Min-Hyuk Seo**, Jeong-Eun Nah***
*Dept. of Applied Statistics, Yonsei University
**Dept. of Computer Science, Yonsei University
***University College, Yonsei University

요 약

COVID-19로 인한 홈 쿠킹 시장 수요 증가로 사람들은 더 편리한 요리 보조 시스템을 필요로 하고 있다. 기존 요리 시스템은 휴대폰, 책을 통해 레시피를 일방적으로 제공하기 때문에 사용자가 요리과정을 중단하고 반복적으로 열람해야 한다는 한계점을 가진다. '대화형 스마트 쿠킹 서비스' 시스템은 요리 과정 전반에서 필요한 내용을 사용자와 상호작용하며 적절하게 인지하고 알려주는 인공지능 시스템이다. Google의 MediaPipe를 사용해 사용자의 관절을 인식하고 모델을 학습시켜 사용자의 요리 동작을 인식하도록 설계했으며, dialogflow를 이용한 챗봇 기능을 통해 필요한 재료, 다음 단계 등의 내용을 실시간으로 제시한다. 또한 실시간 행동 인식으로 요리과정 중 화재, 베임 사고 등의 위험 상황을 감지하여 사용자에게 정보를 전달해줌으로써 사고를 예방한다. 음성인식을 통해 시스템과 사용자 간의 쌍방향적 소통을 가능하게 했고, 음성으로 화면을 제어함으로써 요리과정에서의 불필요한 디스플레이 터치를 방지해 위생적인 요리 환경을 제공한다.

1. 서론

COVID-19 발생 이후 집에서 시간을 보내는 비율이 늘어나면서 홈 쿠킹에 대한 관심도 함께 증가하고 있다. 한국농촌경제연구원(KREI)의 '2020 식품소비행태조사'에 따르면 61.7%가 가정 내 식사 횟수가 증가했다고 답했다. 이에 사용자들은 보다 더 편리하게 레시피 정보를 열람할 수 있는 요리 보조 시스템을 필요로 하고 있다. 기존 요리 환경에서 레시피 정보를 제공하는 Youtube 동영상, 요리 책자, 인터넷 등은 다음 단계의 레시피를 확인하기 위해 요리과정을 중단하고 다시 정보를 열람해야 하는 불편함을 초래하고 있다. 이러한 일방적인 정보 제공의 한계가 있어, 사용자와 상호작용이 가능한 요리 시스템이 요구된다.

이에 본 논문은 사용자가 요리과정을 중단하지 않고 정보를 얻을 수 있는 "대화형 스마트 쿠킹 서비스(An Interactive Smart Cooking Service System)" 시스템을 제안한다. 기존의 요리 시스템이 제공하고 있는 레시피 정보와 더불어, 사용자의 요리 행동을 인식하고, 이를 분석해 레시피 및 화재 등의 위험상황에 대한 정보를 능동적으로 제공한다. 사용자는 요리 과정을 파악하기 위해 요리를 중단해야 하는 상황에서 벗어나 시스템과 상호작용하며 편리하게 요리를 진행할 수 있다. 요리 과정에서 생긴 질문에 대한 답변과 다음단계, 재료 등에 대한 정보를 얻을 수 있으며, 인지하지 못한 위험 상황에 대한 대처가 가능하다. 본 논문에서는 대화형 스마트 쿠킹 서비스 시스템의 설계과정 및 핵심 기능인 행동인식과 음성인식 기술의 구현 결과를 중점적으로 살펴보고자 한다.

2. 관련 연구

초기의 쿠킹 시스템 관련 연구는 행동 인식을 위해 사람의 몸 또는 주변 환경에 sensor를 부착하였다. “Cook’s Collage: Two Exploratory Designs (2001)”[1]의 경우는 사용자가 요리과정 중에 과정이나 물건을 기억하도록 보조하는 것에 초점을 맞추어 상황인식 시스템을 설계하였다. 물건을 추적하기 위해서 감지 센서를 사용하고, 카메라를 사용하였지만 단순한 동작을 캡쳐하여 요리 진행상황을 사용자에게 보여주는 용도로 사용하였다. “Hands On Cooking: Towards an Attentive Kitchen (2003)”[2]은 eyetracking을 위한 sensor와 음성인식을 이용한 시스템을 제안한다. 위의 두 연구의 경우 공통적으로 사용자의 요리 동작을 시스템이 직접 인식할 수 없다는 한계를 가진다.

이후에 카메라를 이용한 동작 인식이 시스템에 접목된다. “Smart kitchen: A user centric cooking support system(2008)”[3]에서는 조리 단계별로 식품과 동작을 인식하는 시스템을 제시하였고, “Cookin’: Interactive Cooking Assistant(2017)”[4]는 동작 인식을 통해서 요리과정에서 시스템과 사용자 간의 상호작용을 시도하였다. 하지만 사용자의 요리과정을 인식하는 것이 아닌 next(다음 과정), again(레시피 다시 보여주기), close(닫기)등과 같이 시스템과의 소통에 관한 정해진 동작만을 인식하고 있다.

‘대화형 스마트 쿠킹 서비스’ 시스템은 명령에 관련한 부분과 사용자와의 의사소통을 speech recognition API를 이용한 음성인식 챗봇으로 설계해 사용자가 명령을 더 다양하고 쉽게 할 수 있도록 하였다. 더 나아가 실제 요리과정을 인식하여 기존 시스템에서의 요리 행동 인식의 한계를 극복했다.

3. 스마트 쿠킹 서비스 시스템 설계

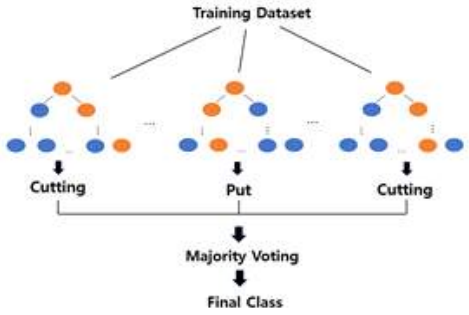
3.1 행동인식 모델

인간 행동 인식 분야에서 딥러닝이 사용된 동영상 분석 기술을 적용해 인간의 다양한 행위를 인식할 수 있게 되었다.[5] 최근 Human Pose Estimation을 위한 딥러닝 기반 키포인트 추출 모델의 연구가 활발히 진행되고 있다. 최초의 실시간 키포인트 검출 시스템인 OpenPose의 등장 이후, 다양한 모델이 탄생하였다.

현재 빈번히 사용되고 있는 모델인 PoseNet의 사용을 고려해 테스트를 진행했다. 하지만 품질과 해상도가 양호해도 피사체가 가까이 있으면 성능이 떨어지는 현상이 발생하였고, 옷의 재질이 신체 키포인트 생성에 영향을 미치는 것을 확인하였다. 본 연구에서 사용되는 디스플레이의 카메라 위치는 피사체와 비교적 가까운 거리에 있어야 했기 때문에, PoseNet의 사용이 어렵다고 판단하였다.

본 연구에서는 Google의 MediaPipe를 사용하였다. MediaPipe는 파이프라인을 구축하여 비디오, 오디오 형식의 데이터를 처리할 수 있는 머신러닝 기반의 “Open Source Crossplatform Framework”이다.[6] 연구에서 사용한 Holistic의 경우 522개의 키포인트를 이용해 행동, 제스처, 얼굴 표정의 전체적 동시 인식을 지원한다.

키포인트 추출 모델로는 BlazePose가 MediaPipe 내에서 사용되었고, 이 모델은 앞서 말한 PoseNet의 단점을 극복한 최신의 모델로 앞서 PoseNet이 한계점을 가졌던 상황에서도 좋은 성능을 보이는 것을 확인하였다. 시스템에서 인식할 썰기, 젓기, 넣기 3가지 행동을 다양한 각도와 장소에서 반복하며, 사용자의 상체 관절 키포인트 33개에 대한 모델 훈련 데이터를 수집하였다. 훈련 결과, (그림 1)과 같은 과정으로 실행되는 Random Forest Classifier에서 가장 높은 정확도를 보였고, 이를 행동 인식 모델로 선정하였다.

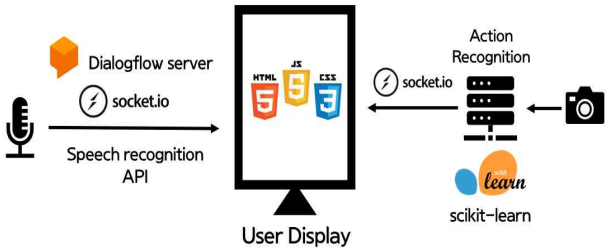


(그림 1) Random Forest Classifier

3.2 행동 및 음성인식 시스템 연결

사용자가 스마트 쿠킹 시스템을 실행하면 미니 PC가 작동하여 LED 모니터에 시작 화면이 출력된다. 사전에 등록된 메뉴 중 하나를 선택하고 해당 메뉴에 해당하는 메인 화면으로 연결되어 요리가 진행된다. 요리를 진행하는 과정에서 음성을 통해 요리과정 전반에 관한 설명을 요청할 수 있으며, 이때 (그림 2)와 같이 Dialogflow server를 통해 적절한 답변이 출력된다. 카메라를 이용한 행동 인식을 통해 요리과정을 전반적으로 점검할 수 있다. 사용자가 요리하는 모습을 실시간으로 인식하여 레시피 순

서대로 요리를 진행하지 않을 때 경고 메시지를 출력하고, 써는 행동이 인식되면 손을 조심하라는 메시지 창을 뜨게 하였다. 사용자의 관절이 5분 이상 인식되지 않았을 때, 위험상황으로 인지하여 경고음을 출력하는 '위험 상황 감지 기능'에도 이용되었다.



(그림 2) 시스템 구성도

Agile 방법론에 따라 요리과정에서 사용자에게 꼭 필요한 기능만을 정리하여 메인 화면에 담았다. 레시피 정보를 텍스트와 영상 형태의 시청각 자료로 제공했고, 요리 과정에 대한 질문과 답변을 채팅 형태의 텍스트로 출력함으로써 원활한 소통을 가능하게 했다. 음성인식을 통해 타이머, 영상 재생 등의 모든 소프트웨어를 동작시켜 사용자가 요리 도중 불필요한 터치를 하지 않도록 설계하였다.

4. 행동 및 음성 인식 시스템 구현

4.1 행동 인식 기능

행동 인식 기능은 크게 '데이터 수집 - 모델 훈련 - 모델 선택'의 과정으로 이루어졌다.

첫 번째, 데이터 수집 단계에서는 인식하고자 하는 3가지 행동 '썰기, 젓기, 넣기'에 대한 각각의 데이터를 수집한다. MediaPipe API를 이용하여 사용자의 신체부위 33곳을 관절 포인트로 인식하고, 이를 (x, y, z, v)로 좌표화한다. 사용자가 3가지 행동을 하는 모습을 1초 단위로 캡처하여 약 17,000개의 데이터를 수집하였고, shape이 (17478, 133)인 데이터 셋을 만들었다.

두 번째, 수집된 데이터를 훈련에 사용하기 위해, Train Data와 Validation Data를 7:3 비율로 나누었다. 어떤 모델이 가장 좋은 검증 정확도를 내는지 비교하기 위해, 확률적 판별모형인 Logistic Regression, 확률적 생성모형인 나이브 베이즈 중 하나인 BernoulliNB, 앙상블 모델인 RandomForest Classifier, GradientBoosting Classifier에 대한 각각의 pipeline을 만들었다.

세 번째, 4개의 모델을 훈련하고 Validation Data를 이용해 검증 정확도를 아래 <표 1>에서 비교하였고, 가장 높은 정확도를 기록한 'RandomForest Classifier'을 최종 모델로 선택하였다.

훈련 모델	정확도(%)
Logistic Regression	98.47
BernoulliNB	84.67
Random Forest Classifier	99.87
Gradient Boosting Classifier	99.68

<표 1> 4개 모델 적합성 검증 정확도 비교

모델의 예측 확률이 0.8보다 큰 경우, (그림 3)에서처럼 인식된 행동의 Class 이름과 그에 대한 확률이 화면에 나오고, 이 외의 경우에는 3가지 행동 외의 행동으로 인식하여 Others class로 분류된다.

(그림 3) 행동 인식 구현 결과

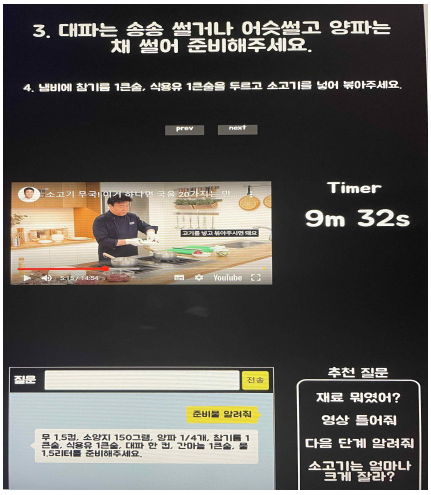
4.2 음성 인식 기능

dialogflow는 사용자의 발화에서 context를 분석하여 사전에 설정한 대답을 제공하는 구글의 챗봇 플랫폼이다. 대화의 흐름을 손쉽게 만들어주는 dialogflow를 통해 요리 진행 흐름, 요리 중의 질문사항 등에 대한 사용자와의 효과적인 상호작용이 가능해진다. 마이크로 입력되는 사용자의 발화를 웹브라우저의 SpeechRecognition(Speech-to-text) API를 통해 텍스트 데이터로 변환한 뒤 node.js를 통해 dialogflow의 서버로 전송한다. dialogflow가 전송받은 사용자 발화를 분석하여 사전에 정의된 적절한 응답을 내보낸다. 서버로부터 받아온 텍스트 형태의 응답을 웹 브라우저의 SpeechSynthesis (Text-to-Speech) API를 통해 음성 데이터로 합성하여 재생함과 동시에 화면에 표시하여 사용자에게 전달한다. 그 외에 dialogflow를 거치지 않아도 되는 단순 조작 명령은 텍스트 데이터로 변환하여 javascript의 조건문을 통해 사용자의 명령을 수행한다.

4.3 시스템 평가

본 연구에서는 실시간으로 사용자의 행동을 인식하여 그에 따른 즉각적인 피드백을 할 수 있는 쌍방

향 의사소통 요리 레시피 제공 시스템을 아래 (그림 4)와 같이 구현하였다. 제공된 레시피에서 해당 단계에 필요한 행동이 One-Hot Encoding 기법을 통해 연동되어 0과 1로 이루어진 벡터로 표현되어 있어 사용자의 요리 순서가 잘못되었을 때 이를 빠르게 바로 잡아줄 수 있다. 또한, 기존에 사용자의 이동을 감지하기 위해 사용했던 PIR 센서 대신 행동 인식의 기능만으로도 사용자의 이동 유무를 감지할 수 있다는 차별점이 존재한다. 이는 기존에 사용자가 어플리케이션이나 인터넷 매체 등에만 의존했던 것과 달리, 요리 도중에 문제가 생겼을 때 챗봇 기능을 통해 즉각적으로 문제에 대한 해답을 얻을 수 있다는 장점이 있다. 더불어 요리 진행 중에 요리와 관련한 정보 확인을 위해 화면을 터치할 필요 없이 인식을 통해 상호작용할 수 있다는 점은 사용자가 느낄 수 있는 편리성을 극대화한다.



(그림 4) 구현된 UI/UX 화면

본 연구의 향후 개발 방안은 다음과 같다. 첫째, 인식되는 행동의 종류를 늘린다. 현재 구현한 시스템에서는 썰기, 젓기, 넣기 총 3가지 행동만을 인식한다. 향후 더 복잡한 레시피를 제공하거나 정밀한 인식을 위해서는 인식 가능한 행동을 세분화하거나 그 종류를 다양하게 할 필요성이 있다. 둘째, 두 명 이상의 사용자가 요리하는 상황에서도 정확하게 인식하도록 한다. 현재는 한 사람의 관절을 좌표화하여 모델을 훈련 시켰기 때문에, 행동 인식 과정에서 카메라에 사용자가 아닌 다른 사람이 인식되면 순간적으로 행동 인식의 정확도가 감소할 수 있다. 요리할 때 다른 사람이 인식될 수 있다는 점을 고려하여 MediaPipe을 이용한 multi hand-tracking 기법을 통해 이를 해결하고자 한다.

5. 결론

최근 홈쿠킹과 이를 보조하는 서비스에 대한 수요가 증가함에 따라, 사용자가 요리과정에서 직접 검색하지 않고도 관련 정보를 쉽게 얻을 수 있도록 쌍방향 소통 레시피 제공 시스템을 개발함으로써 요리 중 발생할 수 있는 불편함을 개선하고자 하였다.

기존 선행 연구에서는 센서, eye-tracking, 음성인식, 카메라 등을 이용하여 요리과정에서 사용자 행동을 피드백하려는 시도가 있었으나, 사용자와의 능동적인 커뮤니케이션을 구현하기에는 한계가 있었다. 기존의 시스템이 갖고 있던 요리과정의 행동 인식 부분의 부족함을 MediaPipe API와 행동 인식 모델링을 통해 비교적 정확한 요리 동작 인식을 하도록 개선하였다. 명령을 interfacing하는 동작이 아닌 실제 요리과정을 인식하도록 하여 기능적으로 더 필요한 부분을 구현하였다. 음성인식 챗봇을 통하여 요리과정을 인식하며 진행되는 쌍방향 소통이라는 점에서 기존의 연구와 차별화하였다.

본 논문은 과학기술정보통신부
정보통신창의인재양성사업의 지원을 통해 수행한
ICT멘토링 프로젝트 결과물입니다

참고문헌

[1] Tran, Q., et al. (2002). Cook's Collage: Two Exploratory Designs. Position paper for Families Workshop at CHI 2002. Minneapolis, MN, 2002

[2] J. S. Bradbury, J. S. Shell, and C. B. Knowles. Hands on cooking: towards an attentive kitchen. Conference on Human Factors in Computing Systems, pages 996 - 997, 2003.

[3] Atsushi HASHIMOTO. et. al, Smart Kitchen: A User Centric Cooking Support System, Proceedings of IPMU'08, Torremolinos (M´alaga), June 22 - 27, 2008, Pages 848 - 854

[4] Urja Khurana. et. al, Cookin': Interactive Cooking Assistant, 2017

[5] Vrigkas, Michalis, Christophoros Nikou, and IoannisA. Kakadiaris, "A review of human activity recognition methods," Frontiers in Robotics and AI,Vol 2, article 28, 2015

[6] Camillo Lugaresi. et al., MediaPipe: A Framework for Building Perception Pipelines, 2019, arXiv:1906.08172