실시간 동영상 스트리밍 환경에서 오디오 및 영상기반 감정인식 프레임워크

방재훈, 임호준, 이승룡
경희대학교 컴퓨터공학과
e-mai: {jhb, lhj, sylee}@oslab.khu.ac.kr

Audio and Image based Emotion Recognition Framework on Real-time Video Streaming

Jaehun Bang, Ho Jun Lim, Sungyoung Lee
Dept of Computer Science and Engineering, Kyung Hee University

요 약

최근 감정인식 기술은 다양한 IoT 센서 디바이스의 등장으로 단일 소스기반의 감정인식 기술 연구에서 멀티모달 센서기반 감정인식 연구로 변화하고 있으며, 특히 오디오와 영상을 이용한 감정인식 기술의 연구가 활발하게 진행되고 있다. 기존의 오디오 및 영상기반 감정인식 연구는 두 개의 센서 데이터를 동시에 입력 저장한 오픈 데이터베이스를 활용하여 다른 이벤트 처리 없이 각각의 데이터에서 특징을 추출하고 하나의 분류기를 통해 감정을 인식한다. 이러한 기법은 사람이 말하지 않는 구간, 얼굴이 보이지 않는 구간의 이벤트 정보처리에 대한 대처가 떨어지고 두 개의 정보를 종합하여 하나의 감정으로 도출하는 디시전 레벨의 퓨저닝 연구가 부족하다. 본 논문에서는 이러한 문제를 해결하기 위해 오디오 및 영상에 내포되어 있는 이벤트 정보를 추출하고 오디오 및 영상 기반의 분리된 인지모듈을 통해 감정들을 인식하며, 도출된 감정들을 시간단위로 통합하여 디시전 퓨전하는 실시간 오디오 및 영상기반의 감정인식 프레임워크를 제안한다.

1. 서론

최근 IoT 기술의 등장으로 생활 속에 분포해 있는 다양한 멀티모달 센서로부터 사용자의 상황 및 의도를 파악하는 인지 기술이 활발하게 연구되고 있다. 감정인식 분야는 사용자의 경험 (UX) 대변하는 중요한 요소로 대화형 시스템, 헬스케어 서비스, 서비스 만족도 평가 등 다양한 산업에서 사용되고 있다.

감정인식 기술은 오디오, 비디오, 텍스트, 뇌파 등 인풋소스에 따라 개별적으로 연구가 진행되어져 왔으나 단일 소스의 활용만으로는 인지 감정의 종류의 한계 및 낮은 정확도의 문제로 최근에는 여러 종류의 인풋을 혼합한 멀티모달 센서 기반의 감정인식 기법에 대한 연구가 진행되고 있다.

그 중 사용자의 음성과 얼굴 이미지를 활용한 오디오 및 영상기반 감정인식분야는 영상통화, 디지털 사이니지 등 다양한 기기에서 손쉽게 구현이 가능하고 활용성이 높아 가장 활발히 연구되고 있는 분야이다.

기존의 오디오 및 영상기반 감정인식 기술은 대부분 eNTERFACE[1], GEMEP[2], IEMOCAP[3] 등 오픈 데이터베이스를 활용하여 높은 정확도 달성을 목표로 하나의 훈련 모델을 사용한 특징추출 알고리즘 연구에 집중되어 있다.[4]

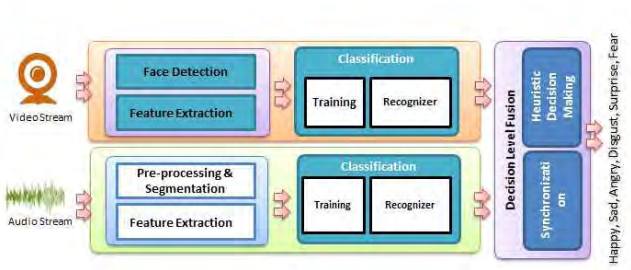
기존 연구들에 활용되었던 데이터베이스는 특정 감정 시연 구간만을 녹화한 동영상 파일로 구성되어 있어, 사실상 동시에 오디오와 영상이 들어오는 인풋 환경 대상으로 연구를 진행해 왔다. 그러나 실제 실시간 환경에서 입력되는 데이터 는 이벤트성의 정보로 오디오의 경우에는 사용자가 말하지 않는 구간이 존재하고 영상의 경우에는 사용자의 얼굴이 보이지 않는 구간에서 정보의 공백이 생길 경우 대처하기 어렵다는 단점이 있다.

이런 문제점들을 해결하기 위해서는 실시간으로 들어오는 동영상 데이터를 영상 및 오디오 데이터로 추출하여 각각의 인지모듈을 통해 감정들을 도출 후 이를 동기화하여 디시전 메이킹을 수행하는 퓨저닝 기술이 필요하다.

따라서 본 논문에서는 지속적으로 입력되는 동영상 스트리밍 환경에서 감정인식 정확도를 높이고 이벤트 정보 처리를 통해 유연한 실시간 감정인식 프레임워크를 제안한다.

2. 실시간 오디오 및 영상기반 감정인식 프레임워크

본 논문에서는 제안하는 실시간 오디오 및 영상기반 감정인식 프레임워크는 음성기반 감정인식, 영상기반 감정인식, 디시전 레벨 퓨전 3단계로 구성되어 있다. 제안하는 프레임워크는 실시간으로 들어오는 동영상정보를 오디오 및 영상 정보로 분리하고 이를 각 해당 소스에 해당하는 인지모듈로 전송하여 감정을 추론 후 디시전 퓨전을 통하여 최종 감정을 도출한다.

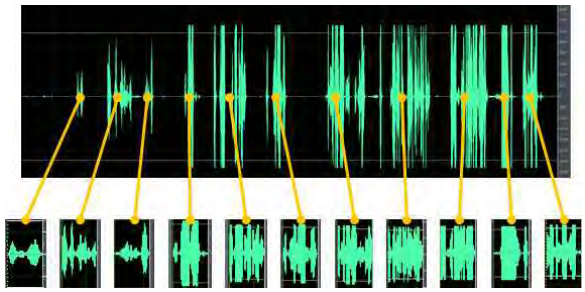


[그림 1] 제안하는 실시간 오디오 및 영상기반 감정인식 프레임워크

2-1. 오디오 기반 감정인식

제안하는 오디오 기반 감정인식은 실시간으로 입력되는 비디오 정보에서 오디오 스트리밍 데이터를 추출하고 사용자가 말하는 음성 구간을 추출하고 해당 구간을 14차 MFCC (Mel Frequency Cepstral Coefficients) [5] 통계적 특징을 추출하여 이를 기반으로 화남, 역겨움, 두려움, 놀람, 슬픔, 행복의 6가지 감정을 인식한다.

사용자의 음성구간을 추출하는 방법은 원 오디오 정보를 데시벨(DB) 정보로 변환 후 일반적으로 속삭이는 음성크기인 15DB을 임계값으로 두어 사용자가 말하지 않는 부분을 제거한다. 추출된 음성정보에서 사용자가 말을 시작한 부분을 Start Point 지정하고 0.5초 이상의 음성공백이 발생하여 추출되는 음성이 없으면 이를 End Point로 규정하여 한 어절의 음성을 추출한다.



[그림 2] 음성구간 추출 예시

이렇게 추출된 한어절의 음성정보를 바탕으로 13차 MFCC 필터뱅크 알고리즘을 사용하여 총 117개의 통계적 특징을 추출하며 구성은 표 1과 같다.

<표 1> MFCC 통계적 특징벡터 목록

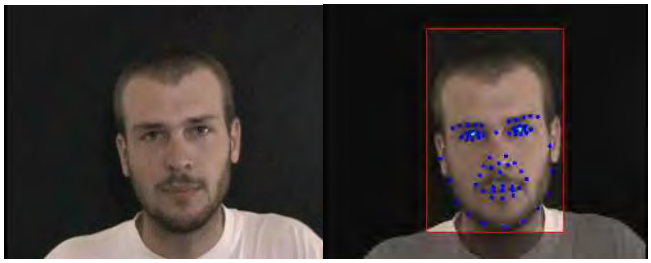
- 13 MFCC Mean Values
- 13 MFCC Standard Deviation Values
- 13 MFCC Maximum Values
- 13 MFCC Minimum Values
- 13 MFCC Rage Values
- 13 MFCC Energy Mean Values
- 13 MFCC Energy Standard Deviation Values
- 13 MFCC Energy Max Values
- 12 MFCC Correlation Values

추출된 특징벡터는 기계학습 알고리즘을 이용하여 훈련 모델을 생성하고 이를 기반으로 총 6가지 감정을 인식한다.

2-2. 영상 기반 감정인식

제안하는 영상기반 감정인식은 입력되는 동영상 스트리밍 정보에서 1초당 하나의 영상 정보를 추출하고 얼굴인지를 통해 사람의 얼굴 검출 및 특징점을 추출하여 이를 기반으로 총 14개의 각도 및 길이 특징 벡터를 추출한 뒤 기계학습 알고리즘을 통해 화남, 역겨움, 두려움, 놀람, 슬픔, 행복의 6가지 감정을 인식한다. [6]

얼굴검출 및 특징점 추출 방법으로는 Luxand Face SDK[7]를 활용하여 총 66개의 얼굴 특징점을 추출한다. 검출된 얼굴 및 특징점들은 그림 3과 같이 표현된다.



[그림 3] Luxand Face SDK를 활용한 얼굴 검출 및 특징점 추출

추출된 66개의 얼굴 특징점들을 기반으로 감정에 영향을 가장 많이 주는 구성요소인 '눈썹, '눈, '입' 3가지의 각도, 거리의 총 14개의 통계적 특징을 추출하며 특징벡터의 구성은 표 2와 그림 4와 같다.

<표 2> 감정얼굴 특징 종류

• 왼쪽 눈썹 각도	• 오른쪽 눈썹 각도
• 왼쪽 눈 길이	• 오른쪽 눈 길이
• 왼쪽 눈 높이	• 오른쪽 눈 높이
• 왼쪽 눈위 각도	• 오른쪽 눈위 각도
• 왼쪽 눈아래 각도	• 오른쪽 눈아래 각도
• 입 길이	• 입 높이
• 위쪽 입술 각도	• 아래쪽 입술 각도

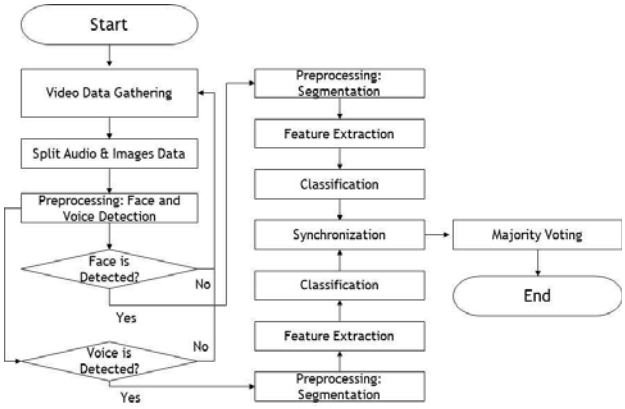
[그림 4] 각도별 얼굴감정특징(a)
거리별 얼굴감정특징(b)

추출된 감정얼굴 특징벡터는 오디오 기반 감정인식과 마찬가지로 기계학습 알고리즘을 이용하여 훈련모델을 생성하고 이를 기반으로 총 6가지 감정을 인식한다.

2-3. 디시전 레벨 퓨전

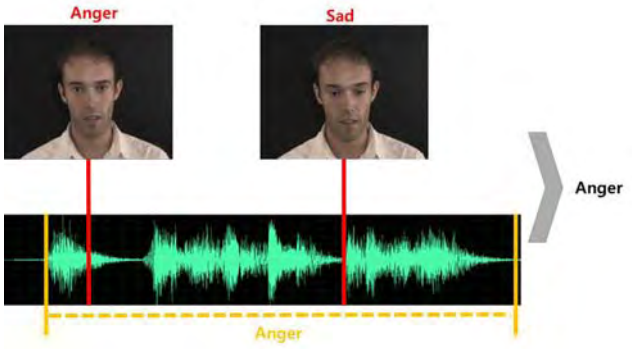
디시전 레벨의 퓨전은 오디오와 영상을 기반으로 도출된 감정 정보들을 동기화 및 취합하여 최종 감정정보를 도출한다. 감정정보 동기화 과정에서 오디오 데이터는 연속적이면서 충분한 사용자 음성 데이터가 축적이 되어야지만 감정인식이 가능하며, 영상의 경우 하나의 이미지에서 사용자의 얼굴이 검출되어야만 감정이 인식되는 불연속 정보로써, 두가지 종류의 데이터가 모두 이벤트성 데이터로써 이를 정확하게 동기화 한다는 것은 어렵다.

따라서 제안하는 디시전 레벨 퓨전 기법은 이러한 이벤트성 오디오 및 영상 데이터를 최소한의 과정을 통해 동기화 하고 해당 구간에 따른 Majority Voting기반의 디시전 메이킹 알고리즘을 통해 최종 감정을 도출한다. 제안하는 디시전 레벨 퓨전 동작의 순서도는 그림 5와 같다.



[그림 5] 디시전 퓨전 동작 흐름 순서도

단일 센서의 이벤트만 감지할 경우 해당 인지 모듈을 통해서만 감정인 인식되며 음성 및 영상이 동시에 들어오는 구간에서는 Majority Voting 통하여 최종감정을 도출하며 Majority Voting의 동작원리는 그림 6와 같다.



[그림 6] 디시전 레벨 퓨전 동작원리

3. 실험 환경 및 결과

본 장에서는 제안하는 프레임워크의 검증을 위하여 오픈 감정 데이터 셋인 eNTERFACE를 사용한다. eNTEFACE 데이터 셋은 시정각 데이터 베이스로 14개의 서로 다른 국가에서 42명의 실험군을 (남자:34명, 여자 8명) 대상으로 서로 다른 감정을 자극시키는 6개의 짧은 구절을 연기를 통해 총 1263개의 동영상을 녹화한 감정인식 분야에서 성능평가를 위해 가장 많이 활용되는 데이터 셋이다.

본 실험은 eNTERFACE 데이터 셋의 동영상을 기반으로 오디오 및 동영상기반 감정인식의 알고리즘

에 대해 여러 분류 알고리즘을 적용하여 10-fold cross validation 기법을 통해 검증한다.

오디오 기반 감정인식의 경우 가장 높은 성능을 보인 분류 알고리즘은 Random Forest 알고리즘으로 평균 80.2%의 정확도를 보였으며 해당 알고리즘의 Confusion Matrix는 표 4와 같다.

<표 3> 분류알고리즘별 평가결과

분류알고리즘	정확도 (%)
Random Forest	80.2
IBK	76.02
KStar	75.25
SVM	70.69
J48	69.45

<표 4> Random Forest 알고리즘 분류 결과

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**84.72222**	2.777778	1.388889	4.166667	1.388889	5.555556
Disgust	3.240741	**79.62963**	4.62963	5.092593	5.555556	1.851852
Fear	7.407407	2.777778	**70.83333**	1.851852	12.96296	4.166667
Happy	5.164319	6.103286	1.877934	**81.22066**	1.877934	3.755869
Sad	0	0.925926	6.018519	0.462963	**88.42593**	4.166667
Surprise	5.092593	4.166667	3.240741	2.777778	8.333333	**76.38889**

영상기반 감정인식의 경우 동영상의 초당 한 장의 이미지를 추출하여 정확도를 검출하므로 표본의 개수가 총 4,610개의 데이터로 실험을 진행하였다. 가장 높은 정확도를 보인 알고리즘은 Random Forest 알고리즘으로 평균 79.93%의 정확도를 보였으며 해당 알고리즘의 Confusion Matrix는 표 6과 같다.

<표 5> 분류알고리즘별 평가결과

분류알고리즘	정확도 (%)
IBK	81.45
Random Forest	79.93
KStar	79.93
J48	67.44
SVM	32.68

<표 6> IBK 알고리즘 분류 결과

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**82.71186**	4.067797	3.615819	4.067797	3.163842	2.372881
Disgust	3.557312	**79.57839**	3.820817	4.479578	4.743083	3.820817
Fear	4.539202	2.61348	**80.46768**	3.576341	6.189821	2.61348
Happy	2.052786	5.131965	2.052786	**84.45748**	2.199413	4.105572
Sad	3.567036	5.166052	5.781058	3.690037	**78.35178**	3.444034
Surprise	6.048387	3.091398	1.747312	2.150538	3.494624	**83.46774**

4. 결론 및 향후 연구

본 논문에서는 이벤트성 데이터로 동기화가 어려운 실시간 비디오 스트리밍 환경에서 오디오 및 영상기반의 감정인식 프레임워크를 제안하고 실제 구현 하였다. 오디오 및 영상 감정인식에 대한 알고리즘 평가를 eNTEFACE 데이터 셋으로 검증하였고 두 개의 인지 모듈이 평균 80%의 정확도를 보였다. 향후 연구로는 다양한 사람들의 연속적인 감정 동영상 정보를 수집하여 제안하는 프레임워크의 복합 정확도를 측정하고 제안한 디시전 퓨전 알고리즘을 고도화할 계획이다.

5. Acknowledgement

This work (Grants No.C0395816) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2016.

참고문헌

[1] Martin, O., Kotsia, I., Macq, B., Pitas, I., “The eNTERFACE’05 audiovisual emotion database”, Int. Conf. Data Engineering Workshops, 2006.

[2] Bänziger, T., Pirker, H., Scherer, K., “Gemep - Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions”, Proc. of LREC Workshop on Corpora for Research on Emotion and Affect, 2006, pp.15 - 19

[3] Paleari, M., Benmokhtar, R., Huet, B., “Evidence theory-based multimodal emotion recognition”, Proc. 15th Int. Multimedia Modeling Conf. Advances in Multimedia Modeling, 2009, pp.435 - 446

[4] Wu, C. H., Lin, J. C., & Wei, W. L., “Survey on audiovisual emotion recognition: databases, features, and data fusion strategies”, APSIPA transactions on signal and information processing, 3, 2014.

[5] A. Klautau "The MFCC", [Online]. Available: http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf

[6] 방재훈, 이승룡. "스마트폰 환경에서 영상기반 실시간 감정인식 프레임워크." 한국정보과학회 학술발표논문집, (2015.06): 443-445.

[7] Luxand Face SDK, (Link: https://www.luxand.com/)