Automatic Multimodal Emotion Recognition Using Facial Expression, Voice, and Text

Hélène Tran[1,2]
[1]Université Clermont-Auvergne, CNRS, Mines de Saint-Étienne,
Clermont-Auvergne-INP, LIMOS, 63000 Clermont-Ferrand, France
[2]Jeolis Solutions, 63000 Clermont-Ferrand, France
helene.tran@doctorant.uca.fr

Abstract

It has been a long-time dream for humans to interact with a machine as we would with a person, in a way that it understands us, advises us, and looks after us with no human supervision. Despite being efficient on logical reasoning, current advanced systems lack empathy and user understanding. Estimating the user's emotion could greatly help the machine to identify the user's needs and adapt its behaviour accordingly. This research project aims to develop an automatic emotion recognition system based on facial expression, voice, and words. We expect to address the challenges related to multimodality, data complexity, and emotion representation.

1 Motivation

Neuroscience research has demonstrated the crucial role of emotions on human decision making [Goleman, 1995]. Therefore, advanced systems involving human users are better off integrating emotions in the loop. Conversational agents, recommendation systems, and virtual reality games are examples of such systems. Another application is to improve remote patient education by customizing physical activity coaching (e.g., to fight obesity). It is to assess user motivation and adapt the exercises accordingly that we aim to develop an automatic emotion recognition system.

Emotions are expressed in different ways (modalities) like facial expression, voice intonation, and body posture. However, many research papers have focused on a single modality. This is not completely reliable as different modalities can give irrelevant or conflicting information on the emotion experienced (cf. figure 1). Multimodal emotion recognition has gained strong interest in the recent years, as it has been shown to improve performance [D'mello and Kory, 2015].

2 Research Project

The objective of the research project is to develop an automatic emotion recognition system based on facial expression, voice, and text inputs. We target videos showing a single person with the face in front of the camera and English or French as spoken language. The naturalistic (or *in-the-wild*) emotions are preferred over acted emotions.

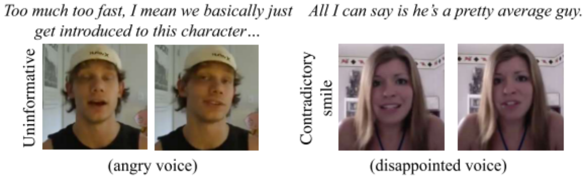
Too much too fast, I mean we basically just get introduced to this character…
All I can say is he's a pretty average guy.
Uninformative
Contradictory smile
(angry voice)
(disappointed voice)

Figure 1: Two data examples from CMU-MOSEI database [Zadeh *et al.*, 2018]

Challenges associated with the project are three-fold:

- **Multimodality.** Fuse modalities which may contain conflicting or redundant information, while taking the context into account. Some questions arise: how to learn interaction within one modality over time while capturing interaction across modalities? Which criteria for identifying relevant features or modalities?
- **Data complexity.** As the three modalities have their own structure, the input data is heterogeneous. To this is added the temporal dimension, which confers a high dimensionality to data. Data are also subject to noise. How to identify noisy sources? Can we preprocess them? Or should we ignore them? For example, Mittal *et al.* [2020] replace the noisy modality by a proxy feature vector calculated from the other modalities.
- **Emotion representation.** There are two main emotional models: discrete or continuous. In the discrete approach, emotions are defined by discrete affective states (e.g., anger, sadness). In the continuous approach, emotions are placed in a continuous space where the main dimensions are valence (pleasant or not?) and arousal (agitated? calm?). Each one has its own advantage: discrete emotions are understood by anybody, while the continuous space can describe a wide range of emotions. The discrete model is chosen for the rest of our work.

Following this analysis, we aim to develop an algorithm for discrete emotion recognition able to:

- extract relevant features and manage conflicting ones
- perform dimensionality reduction and multimodal fusion of *in-the-wild* emotion features
- predict an emotional representation that is both close to the reality and easily described by mathematical tools

3 Contribution

Most of discrete emotion recognition systems only predict one emotion among a predefined list [Zadeh *et al.*, 2018; Tsai *et al.*, 2019; Mittal *et al.*, 2020; Dai *et al.*, 2021]. However, more than experiencing a wide range of emotions, humans often find difficulty in identifying others' emotions with confidence: this is emotional ambiguity.

As databases are the building blocks for the development of such systems, we studied the most popular ones and investigated their position on emotional ambiguity [Tran *et al.*, 2022]. Note that we only focus on multimodal databases which meet our criteria outlined in section 2.

Regarding databases with discrete emotions, 3 out of 5 annotates only one emotion per sentence. As for the CMU-MOSEI [Zadeh *et al.*, 2018] and CMU-MOSEAS [Zadeh *et al.*, 2020] databases, they provide an emotion profile where they assign a level of presence for each of the six primary emotions (anger, disgust, fear, happiness, sadness, surprise).

4 Perspectives

Some works have attempted to represent ambiguity in emotional models: for instance, the winners of Challenge-HML 2018 [1] and 2020 [2], organized by the designers of CMU-MOSEI, used an early fusion network to estimate the presence score of each emotion [Williams *et al.*, 2018] and transformer encoding to perform multi-label classification [Delbrouck *et al.*, 2020] . Since the consideration of ambiguity is relatively new, there is still room for improvement for emotion recognition systems.

As a result, we aim to design a trimodal emotion recognition model able to recognize emotions with their ambiguity. The CMU-MOSEI and CMU-MOSEAS databases seem suitable for our task: in addition to meeting our criteria and introducing emotional ambiguity, the videos show many different people with naturalistic emotions and various topics, which is beneficial for model robustness. Moreover, a large community in affective research trains their model on the CMU-MOSEI database, thus providing reliable benchmarks to compare the effectiveness of our forthcoming architecture.

Acknowledgements

This work is currently being done under a CIFRE contract between the LIMOS laboratory and the Jeolis Solutions company, and sponsored by National Association for Research and Technology (ANRT). It is supervised by Dr. Issam FALIH (issam.falih@uca.fr), Dr. Xavier GOBLET (xavier.goblet@lojelis.com), and Prof. Engelbert MEPHU NGUIFO (engelbert.mephu_nguifo@uca.fr).

References

[Dai *et al.*, 2021] Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. Multimodal End-to-End Sparse Model for Emotion Recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, 2021.

[Delbrouck *et al.*, 2020] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, 2020.

[D'mello and Kory, 2015] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

[Goleman, 1995] Daniel Goleman. *Emotional Intelligence*. Bantam Books, 1995.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1359–1367, 2020.

[Tran *et al.*, 2022] Hélène Tran, Lisa Brelet, Issam Falih, Xavier Goblet, and Engelbert Mephu Nguifo. L'ambiguïté dans la représentation des émotions : état de l'art des bases de données multimodales. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-38:87–98, 2022.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Williams *et al.*, 2018] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, 2018.

[Zadeh *et al.*, 2018] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[Zadeh *et al.*, 2020] Amir Zadeh, Yan Sheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1801. NIH Public Access, 2020.

[1] http://multicomp.cs.cmu.edu/acl2018multimodalworkshop-2/

[2] http://multicomp.cs.cmu.edu/acl2020multimodalworkshop/