ELEC-E5510: Speech Recognition

Literature Study Report: Multi-modal Emotion Recognition

Cuong Nguyen and Daniel Campillejo

November 1, 2024

Computers, or machines in general, are becoming more and more integrated into our daily lives. They are used in various applications, such as in healthcare, education, and entertainment. In many of these applications, it is important for the machine to be able to understand the emotional state of the user. For example, in healthcare, a machine that can recognize the emotional state of a patient can provide better care and support. In education, a machine that can recognize the sentiment of a student can provide personalized learning experiences. In other words, emotion recognition in real time is essential for human-computer interaction. Many studies have been conducted to develop systems that can recognize human emotions from data collected from various sources, such as facial expressions, speech, text and physiological signals. Among them, physiological signals are more objective and reliable than other sources [1]. Human emotions can be classified into six fundemental expressions: happiness, sadness, fear, anger, disgust, and surprise. Other variant emotions can be derived from these six basic emotions [3].

An unimodal emotion recognition system uses data from a single source from the aforementioned sources to recognize emotions. However, unimodal emotion recognition systems have limitations, as they rely on a single source of emotional expression, which may not be sufficient to capture the complexity of human emotions. For example, facial expressions alone may not be enough to accurately recognize emotions, as they can be easily influenced by external factors, such as lighting conditions and occlusions. In order to improve the accuracy of emotion recognition, researchers have started to explore the use of multimodal data, which combine data from multiple sources. It has been shown that multimodal emotion recognition is superior to unimodal emotion recognition, as it can capture more information about the emotional expression [2, 1]. Due to the simplicity and efficiency of deep learning models, they have been widely used in multimodal emotion recognition systems [2, 1]. They especially excel in high-level feature extraction for facial images and text. According to [2], there are three fundemental multimodal fusion methods for emotion recogintion:

- Feature-level fusion: Features from different modalities are concatenated and fed into a classifier.

- Decision-level fusion: Classifiers are trained separately for each modality, and the outputs are combined using a fusion rule.
- Hybrid fusion: A combination of feature-level and decision-level fusion.

Feature-level fusion is the most commonly used method in deep learning-based multimodal emotion recognition systems due to ability to integrate rich, complementary information from multiple modalities, enhancing model accuracy by capturing subtle cross-modal emotional cues early in the processing pipeline [2].

References

[1] K. Ezzameli and H. Mahersia. "Emotion recognition from unimodal to multimodal analysis: A review". In: *Information Fusion* 99 (2023), p. 101847. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101847. URL: https://www.sciencedirect.com/science/article/pii/S156625352300163X.

[2] Bei Pan et al. "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods". In: *Neurocomputing* 561 (2023), p. 126866. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126866. URL: https://www.sciencedirect.com/science/article/pii/S092523122300989X.

[3] Jianhua Zhang et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59 (2020), pp. 103–126. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.01.011. URL: https://www.sciencedirect.com/science/article/pii/S1566253519302532.