Proceedings of the workshop on People in Vision, Language, and the Mind @LREC2022 , pages 6–11
Marseille, 20 June 2022

Do Multimodal Emotion Recognition Models Tackle Ambiguity?

Hélène Tran[1,2], Issam Falih[1], Xavier Goblet[2], Engelbert Mephu Nguifo[1]
[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, {issam.falih, engelbert.mephu_nguifo}@uca.fr
xavier.goblet@lojelis.com

Abstract

Most databases used for emotion recognition assign a single emotion to data samples. This does not match with the complex nature of emotions: we can feel a wide range of emotions throughout our lives with varying degrees of intensity. We may even experience multiple emotions at once. Furthermore, each person physically expresses emotions differently, which makes emotion recognition even more challenging: we call this emotional ambiguity. This paper investigates the problem as a review of ambiguity in multimodal emotion recognition models. To lay the groundwork, the main representations of emotions along with solutions for incorporating ambiguity are described, followed by a brief overview of ambiguity representation in multimodal databases. Thereafter, only models trained on a database that incorporates ambiguity have been studied in this paper. We conclude that although databases provide annotations with ambiguity, most of these models do not fully exploit them, showing that there is still room for improvement in multimodal emotion recognition systems.

Keywords: Multimodal learning, Emotion recognition, Ambiguity

1. Introduction

Emotions have always played a fundamental role in human decision making, from choosing what to eat for lunch to choosing a professional career path. Identifying our emotions, understanding why we are experiencing them, and how to act accordingly are essential to our well-being: this is emotional intelligence. Therefore, support systems for patient education must be able to identify user emotion in order to offer tailored content and maintain user motivation in the long term. Emotion recognition can benefit various other applications such as remote patient follow-up, recommendation systems, and gaming experience.

The development of emotion recognition systems comes with its own challenges. First, many researchers recommend combining multiple sources of information (e.g., voice, text, facial expression) to perform emotion recognition. This is not surprising given the multimodal nature of emotional expression and the human ability to manipulate facial expression or spoken words. Second, the identification, expression, and recognition of emotions can sometimes be tricky, due to the ambiguous nature of emotions. Ambiguity and uncertainty, although closely related, are two distinct ideas: while uncertainty refers to what is not certain to be observed, ambiguity refers to an equivocal trait, where the observed emotion may be confusing. For instance, anger and disgust are two emotions with similar facial expression features. Observing a slightly raised corner of the lip can be open to interpretation (e.g., sarcasm, satisfaction). Emotional ambiguity also includes the observation of several emotions: for example, anger is often mixed with sadness. As a result, databases and machine learning models should consider ambiguity in emotion representation to match what is observed in real life and thus developing more accurate models.

Given the two above challenges, our main objective is to implement a multimodal emotion recognition system based on facial expression, voice, and text data, while taking ambiguity into account. To this end, the paper offers a review of ambiguity in multimodal emotion recognition models by reporting the emotional representation produced in the model output.

The rest of the paper is divided as follows: section 2 presents the two main neural architectures used for model categorisation in the review. Section 3 describes the current emotion representations in the literature and how ambiguity can be incorporated. Section 4 gives a brief overview of multimodal databases that attempt to represent ambiguity, while section 5 is a review of multimodal emotion recognition models with a study of emotion representations in the output. Section 6 discusses their position regarding emotion ambiguity and section 7 concludes the paper with future works.

2. Background

This section describes the main neural architectures involved in the models of our review presented in section 5: recurrent neural networks and transformers.

2.1. Recurrent Neural Networks

Considering the time dimension is relevant when working with sequences. Recurrent neural networks (RNN) are a sub-family of neural architectures designed to operate on temporal sequences. They are equipped with

memory cells to save internal states while processing temporal data sequentially. The most popular RNNs are bidirectional, long short-term memory (LSTM) and gated recurrent units (GRU).
Bidirectional RNNs, presented by Schuster and Paliwal (1997), are composed of two hidden layers which read the input sequence in the forward and backward direction respectively. LSTM and GRU networks, introduced by Hochreiter and Schmidhuber (1997) and Cho et al. (2014) respectively, intend to mitigate the vanishing gradient that traditional RNNs regularly face. The vanishing gradient happens during backpropagation when the gradient becomes smaller and smaller as we come close to the earliest timepoints, until there is no weight update; in this case, the effects of earlier inputs are not learned anymore. LSTM and GRU have similar architecture, with fewer parameters for GRU.

2.2. Transformers

Vaswani et al. (2017) presented a groundbreaking network that has quickly become the basis of numerous deep learning models: transformers. This architecture is an encoder-decoder system that transforms one sequence into another. Transformers rely on an attention mechanism: they identify parts of the sequence representing key information and assign them a higher weight. Since they process sequences as a whole, transformers show better performance than RNNs which rely on long-term dependency and thus face the problem of vanishing gradients. Transformers were originally designed to perform translation tasks and are now widely used in natural language processing.

3. Current Emotion Representations

This section gives an overview of the current emotion representations found in the literature. Subsection 3.1 describes the two main emotional models: discrete and continuous. Subsection 3.2 presents the main limitation of current emotional representations while subsection 3.3 depicts approaches to incorporate ambiguity.

3.1. Main Emotion Representations

The two main representations of emotions are:

- **Discrete.** Emotions are represented by discrete affective states. The most popular list of emotions used in affective computing is that of Ekman (1992): anger, disgust, fear, joy, sadness, and surprise. Another common discrete emotional model is the Wheel of Emotions proposed by Plutchik (2001) which comprises of four pairs of opposite emotions (joy and sadness, trust and disgust, fear and anger, anticipation and surprise) with four degrees of intensity for each emotion (figure 1).

- **Continuous.** Emotions are placed in a multidimensional space. The two main dimensions are *valence* (pleasantness) and *arousal* (measure of physiological activity felt). A third dimension can be added such as *dominance* (Russell and Mehrabian, 1977), which refers to one's ability to take action on the situation, or *potency* (Schlosberg, 1954) which estimates the attention or rejection level towards an object, person, or situation.

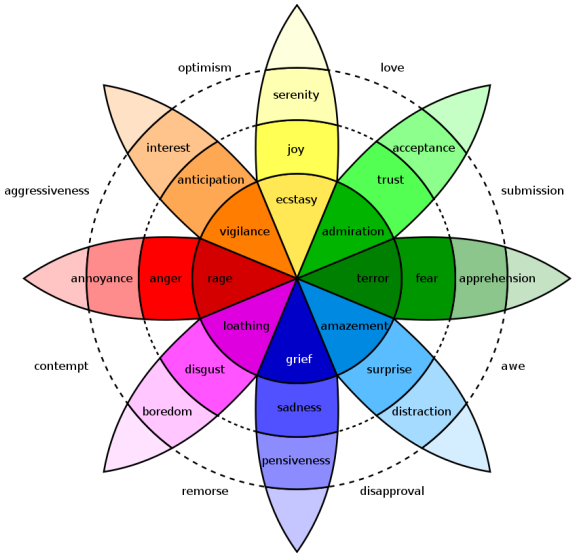

Figure 1: Plutchik's Wheel of Emotions

3.2. Limitation of Current Emotion Representations

Emotions are often represented as a single point. In the discrete approach, only one emotion can be recognized in each sample. In the continuous approach, a single point representing the emotion moves over time in the multidimensional space.
Choosing a punctual representation means being certain about the nature of the emotion perceived. The inherent ambiguity of emotions is not considered here, which might have a negative impact on the accuracy of emotion recognition systems. Gref et al. (2022) analyzed the influence of the ambiguity brought by human annotation in the performance of machine learning models. In their experiments, annotators often combine emotions that are not among the predefined list (e.g., fear and sadness leading to helplessness). This supports their assumption that choosing among the six emotions of Ekman (1992) is not enough to model emotion complexity and that machine learning systems might fail at recognizing the right emotion. Since these results were obtained from a separate analysis of the visual, vocal and textual modalities, a multimodal fusion could perhaps mitigate the ambiguity brought by emotions, hence the motivation for our study.

3.3. Integration of Emotional Ambiguity

There is a growing interest regarding the problem of emotional ambiguity in the affective research community. Some researchers address this issue when implementing their emotion recognition systems (Kim and

Kim, 2018; Fujioka et al., 2020; Li et al., 2021). Sethu et al. (2019) conducted a comprehensive study on introducing ambiguity in the representation of emotions. A summary of the main methods is proposed here.

3.3.1. Discrete Emotions

A second underlying emotion can be identified to complete information on the observed emotion. Vidrascu and Devillers (2005) propose to use major and minor emotions. By extension, an emotional profile can be established where the level of presence of each primary emotion is estimated (Mower et al., 2010). This is a potential solution to the problem outlined by Gref et al. (2022) (cf. section 3.2).

3.3.2. Continuous Emotions

The emotion can be represented using a Gaussian distribution instead of a point (Han et al., 2017): each data sample is associated with the mean and standard deviation of this distribution. Dang et al. (2017) propose not to be restricted to the Gaussian distribution by using a Gaussian mixture model.

4. Multimodal Databases and Representation of Emotion Ambiguity

Databases are the building blocks of the development of emotion recognition systems. Therefore, the choice of the database used for experiments must be thoughtful. If the annotation method does not consider emotional ambiguity, then machine learning models trained on these data will not take it into account either.

Tran et al. (2022) offer a review of multimodal databases with a study of emotion ambiguity in data annotations. They focus on databases which contain facial expression, voice, and text and with English or French as language of speech. They found that among eight reported databases, only CMU-MOSEI (Zadeh et al., 2018b) and CMU-MOSEAS (Zadeh et al., 2020) attempt to represent emotional ambiguity. Both datasets have chosen a discrete model: each data sample is associated with an emotional profile, where a score from 0 to 3 describes the level of presence for each of the six emotions of Ekman (1992). The next section focuses on a review of emotion recognition models trained on CMU-MOSEI (figure 2), a key database in multimodal affective research.

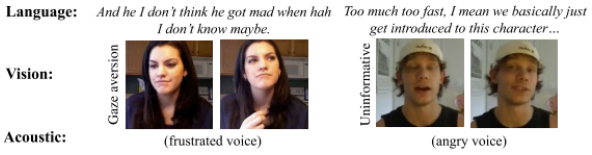

Figure 2: Examples extracted from CMU-MOSEI database (Zadeh et al., 2018b)

5. A Review of Multimodal Emotion Recognition Models

Once a database is chosen, the next step is to design a machine learning model capable of processing annotated data that consider ambiguity, training on them, and recognizing an ambiguous representation of an emotion. In the following, we will focus on the evaluation of the last aspect: the output of the model.

Our review of the multimodal emotion recognition models is comprised of eleven architectures trained on CMU-MOSEI database. All models will be described in subsections 5.1, 5.2, and 5.3. Subsection 5.4 concludes the section with a study of the emotional representations recognized by the models.

5.1. Recurrent Neural Networks

The models falling into this category use either bidirectional, GRU, or LSTM layers (cf. section 2.1). Some perform classification by predicting one or many emotions, others estimate the presence score for each.

5.1.1. Predicting One or More Emotions

Multilogue-Net (Shenoy and Sardana, 2020) is the only reported RNN to predict only one emotion. It uses GRU layers to capture the conversation context and record previous states and emotions while modeling the dependency between interlocutors.

Graph-MFN (Zadeh et al., 2018b) and M3ER (Mittal et al., 2020) both perform binary classification for each emotion. Graph-MFN encodes the three modalities with LSTM layers and uses an interpretable fusion graph to feed its multimodal state memory. This one records the history of interactions between modalities over time. M3ER intends to be robust to noise: it replaces noisy modalities with proxy vectors calculated from the other modalities. Multimodal fusion is done using Memory Fusion Network (Zadeh et al., 2018a), a model with the same architecture as Graph-MFN but with a different fusion module.

5.1.2. Estimating the Presence Score

The two models of this subsection are designed to estimate the intensity of each emotion, rather than detecting the presence of each. The one proposed by Beard et al. (2018) aims to improve Graph-MFN by revisiting the cell memory history of input data encoding layers several times and thus capturing multimodal interactions in the best possible way. With a model training based on L2 loss, their best weighted accuracy is 61.6%.

Williams et al. (2018) attempt to estimate the score of presence with their network composed of bidirectional LSTM layers. Their model is based on early fusion: this means that vectors from audio, image, and text are concatenated before any operation. They perform a custom split 76/14/10 and use a mean absolute error as loss function to select the best model. They obtained a mean unweighted accuracy of 90.6% on the test set.

5.2. Transformer-Based Models

The models of this category use transformers for each modality to extract features. All are designed to predict many emotions (multi-label classification).

MulT (Tsai et al., 2019) is a multimodal fusion model which leverages the benefits of transformers to process unaligned sequences. In the transformer-based joint encoding (TBJE) model by Delbrouck et al. (2020), every modality is encoded jointly before being fed into its respective transformer. Dai et al. (2021) implement a multimodal fusion model able to recognize the emotion directly from raw data. As this can quickly lead to computational overload, an alternative model which inputs the relevant regions of interest extracted from raw data has been developed by the same authors.

5.3. Other Models

Two models using a different architecture are proposed by Lee et al. (2018) and Dai et al. (2020). Both perform classification tasks, the former predicting one emotion and the latter multiple emotions.

Lee et al. (2018) perform multimodal fusion by computing an attention matrix which is the dot product of vocal and textual feature vectors. Their model is composed of three convolutional neural networks: two for vocal and textual feature extraction and one after the attention matrix for the final classification.

Dai et al. (2020) aim to meet the challenges related to unseen or rarely experienced emotions. They built three emotional embedding spaces (textual, visual, and acoustic). Two functions map emotional word embeddings into visual and acoustic spaces. This process can be done for both input data and emotion classes. The final classification is based on the distance between the input sequence and the target emotions. A threshold is set to decide the presence of each emotion.

5.4. Recognizing Emotional Ambiguity

Analyzing the output of an emotion recognition system is a way to study how ambiguity is considered. Out of eleven models, nine consider emotional ambiguity: seven perform multi-label classification and two attempt to estimate the emotion intensity by predicting its presence score. These two models are that of Beard et al. (2018), which attempts to improve Graph-MFN by revisiting the history of cell memories, and the early fusion network of Williams et al. (2018). Since these are recurrent neural networks, they use an activation function that continuously maps to a range of values (e.g., linear, sigmoid) for each output neuron to estimate the presence score of each emotion.

The papers of Dai et al. (2021) and Delbrouck et al. (2020) put together offer a comparison of six out of seven reported models doing multi-label classification: all show similar performance in each of the articles. Unfortunately, we did not find any comparative table of results that involves at least one of the two models which estimate the emotional profile.

6. Discussion

A review of multimodal fusion models for emotion recognition is conducted with a focus on their output. In the case of discrete emotion representation, not considering emotion ambiguity means predicting only one emotion. Two ways to introduce ambiguity would be to predict many emotions and to assess the presence of each emotion (emotional profile). This leads to two different tasks: the former is multi-label classification while the latter is regression for each emotion.

It would have been of interest to compare two models which perform different tasks (predicting one emotion, predicting multiple emotions, or assessing the presence score of each emotion), yet the metrics are not comparable as they all involve different problems.

The main point is that annotations proposed by CMU-MOSEI are not yet fully exploited: many models still perform classification by identifying solely the emotions present in the sample. Therefore, further efforts are needed to assess the intensity of each emotion.

7. Conclusion and Future Work

Developing a multimodal emotion recognition system can be very challenging because of emotion ambiguity arising from human annotation. This can be especially true in a context where many subtle emotions are experienced at the same time in an uncontrolled setting. Emotion ambiguity must first be considered at the level of data annotations and second at every stage of the development of machine learning models, from data preprocessing and model training to final classification.

Among multimodal fusion models trained on a dataset that introduces emotional ambiguity, most perform multi-label classification while a few try to assess the intensity of each emotion. In the next step of our research, we plan to design an emotion recognition system that performs multimodal fusion from visual, vocal, and textual data and is capable of predicting the presence score of each emotion class. The training will be on CMU-MOSEI, a key database for multimodal emotion recognition. Another interesting work would be to analyze the impact of considering ambiguity on the model performance. For instance, there are two ways to address the problem of predicting many emotions: the first by estimating the presence score and setting a threshold to decide which emotions are present and the second by performing binary classification per class (ambiguity less considered than the former).

8. Acknowledgments

This work is done under a CIFRE contract between LIMOS laboratory and Jeolis Solutions company. We thank the National Research and Technology Association (ANRT) for its funding, Ms. Baraa Mohamad and Ms. Lisa Brelet for taking part in the discussions, Mr. Sk. Imran Hossain for English proofreading, and all the members of LIMOS Miners research team for their support.

9. Bibliographical References

Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Eerens, L., Swietojanski, P., and Miksik, O. (2018). Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 251–259.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Dai, W., Liu, Z., Yu, T., and Fung, P. (2020). Modality-Transferable Emotion Embeddings for Low-Resource Multimodal Emotion Recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 269–280.

Dai, W., Cahyawijaya, S., Liu, Z., and Fung, P. (2021). Multimodal End-to-End Sparse Model for Emotion Recognition. pages 5305–5316.

Dang, T., Sethu, V., Epps, J., and Ambikairajah, E. (2017). An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression. In *INTERSPEECH*, pages 1248–1252.

Delbrouck, J.-B., Tits, N., Brousmiche, M., and Dupont, S. (2020). 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.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

Fujioka, T., Homma, T., and Nagamatsu, K. (2020). Meta-Learning for Speech Emotion Recognition Considering Ambiguity of Emotion Labels. In *INTERSPEECH*, pages 2332–2336.

Gref, M., Matthiesen, N., Venugopala, S. H., Satheesh, S., Vijayananth, A., Ha, D. B., Behnke, S., and Köhler, J. (2022). A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis. *arXiv preprint arXiv:2201.06868*.

Han, J., Zhang, Z., Schmitt, M., Pantic, M., and Schuller, B. (2017). From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Kim, Y. and Kim, J. (2018). Human-Like Emotion Recognition: Multi-Label Learning from Noisy Labeled Audio-Visual Expressive Speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5104–5108. IEEE.

Lee, C. W., Song, K. Y., Jeong, J., and Choi, W. Y. (2018). Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data. In *Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 28–34.

Li, Z., Xie, H., Cheng, G., and Li, Q. (2021). Word-level Emotion Distribution with Two Schemas for Short Text Emotion Classification. *Knowledge-Based Systems*, 227:107163.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 1359–1367.

Mower, E., Matarić, M. J., and Narayanan, S. (2010). A Framework for Automatic Human Emotion Classification Using Emotion Profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070.

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11(3):273–294.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, 61(2):81.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., and Narayanan, S. (2019). The Ambiguous World of Emotion Representation. *arXiv preprint arXiv:1909.00360*.

Shenoy, A. and Sardana, A. (2020). Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28.

Tran, H., Brelet, L., Falih, I., Goblet, X., and Mephu Nguifo, E. (2022). 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.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,

Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Vidrascu, L. and Devillers, L. (2005). Real-life Emotion Representation and Detection in Call Centers Data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 739–746. Springer.

Williams, J., Kleinegesse, S., Comanescu, R., and Radu, O. (2018). Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In *Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19.

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., and Morency, L.-P. (2018a). Memory Fusion Network for Multi-View Sequential Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zadeh, A. B., Liang, P. P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., Chen, M., and Morency, L.-P. (2018b). 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*, pages 2236–2246.

Zadeh, A. B., Cao, Y. S., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1812.