






Evidence-Based Research on Multimodal Fusion Emotion Recognition

Zhiqiang Huang  and Mingchao Liao  

School of Mathematics and Computer Science, Wuhan Polytechnic University,
WuhanHubei 430048, China
411781160@qq.com

Abstract. Multimodal fusion classifications are more generalizable and may be utilized in a variety of domains, including medical care, automotive autopilot, and in this paper's study of sentiment identification. This study is motivated by the human perception technique for emotion; it merges the information from auditory and visual modalities to create a novel multimodal fusion emotion algorithm; and it conducts tests to confirm the algorithm's stability. The uncertainty is used as fuzzy propositions for further decision fusion by quantitative calculation of uncertainty, and a credible identification choice is ultimately generated by merging D-S evidence theory. The suggested fusion approach achieves 81.25 percent identification accuracy on the MELD dataset.

Keywords: Multimodal fusion · uncertainty · D-S evidence theory · MELD

1 Introduction

D'Mello et al. conducted an experiment to compare the accuracy of unimodal and multimodal expression identification using a statistical method; the results showed that multimodal performance was experimentally superior than unimodal performance. The datasets that were used in the experiment included a variety of different types. There is a possibility that the researcher known as Chen M. incorporated all of this data in the paper that they wrote (2018). The McGurk phenomenon demonstrates that when the brain perceives, the various senses are spontaneously and instinctually merged to interpret information, and any sensory input is processed as a whole. This is demonstrated by the fact that when the brain perceives, it processes information as a whole. The fact that the brain analyzes information in its entirety as it senses is evidence of this proposition. This is shown by the fact that when the brain processes information, it does it in its whole while it is experiencing it at the same time. This process takes place whenever the brain is presented with new information for the very first time.

In the case that the brain gets inadequate or faulty sensory input, the brain will interpret information received from the outside world with a skewed point of view. Because of this, the brain is more likely to be open to the possibility of forming incorrect conclusions. As this is the case, multimodal feature fusion identification technology has been a popular topic of study over the course of the past two years (D' MelloS K,

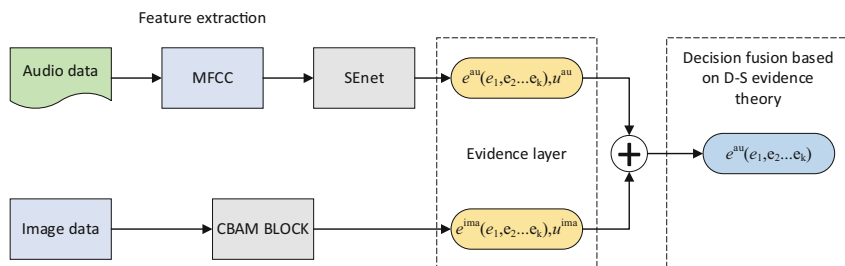


Fig. 1. Evidence Theory Multimodal Fusion Framework Diagram

2015; Noda K, 2014). This is due to the fact that this particular circumstance exists. In connection with the matter at hand, China has attained a certain level of accomplishment up to the current day. Han, Zhang, and their coworkers introduced a new paradigm for multi-view learning with their plausible multi-view classification technique (Srivastava N, 2014) for multi-view learning by dynamically integrating diverse viewpoints at the evidence layer. This new paradigm was based on the idea that multi-view learning can be improved through the integration of multiple perspectives. This new paradigm was founded on the premise that multi-view learning may be enhanced by the inclusion of many views. This thought was the impetus for the development of this new paradigm. This new paradigm is based on the idea that simultaneously taking into account more than one perspective may make multi-view learning more successful. This idea was the impetus for the development of this new paradigm. The purpose of doing this was to make the process of collecting information from a variety of perspectives more straightforward. Classification.

2 Method

2.1 Model Framework

This paper's multimodal fusion method belongs to the late fusion algorithm, also known as the decision layer-based multimodal fusion technique. This chapter presents an uncertainty-based multimodal emotion decision fusion approach to further enhance the model recognition by fusing the first two modal classification models at the decision level. The framework model diagram appears as follows (Fig. 1):

2.2 Uncertainty Calculation

Certain basic fusion algorithms are prone to making erroneous fusion judgments during the process of multimodal decision fusion. In this research, the findings produced from the spoken emotion recognition model and the picture emotion recognition model are combined. If the two arrive at the same conclusion, the final classification result has a high likelihood of being accurate. However, if the two modal recognition models generate contradictory findings, the simple decision summing is likely to provide incorrect results, as seen in Fig. 2.

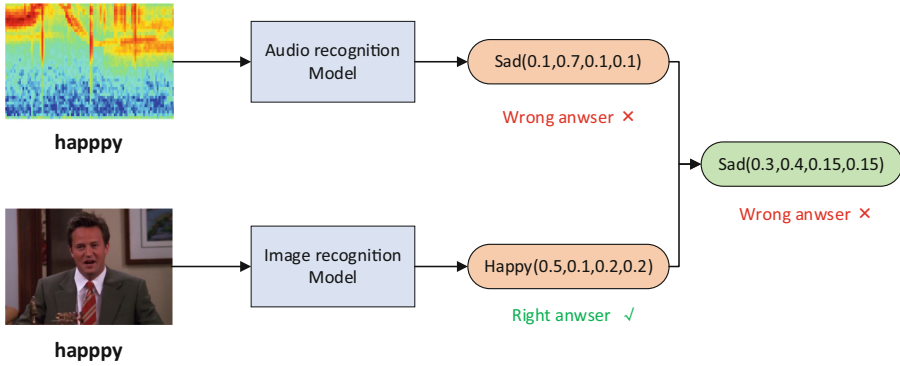


Fig. 2. SoftMax misleading answer graph

As seen in Fig. 2, the model employs the Softmax activation function for the classification job, where the final vector of probability distributions is produced regardless of whether the input data is recognized or not, and the sum of probabilities for each category is 1. In actuality, the model is uncertain that the probabilities should decrease while classifying, hence multimodal input judgments are often overconfident throughout the decision fusion process. Consequently, credible fusion requires a quantitative estimate of the certainty and unpredictability of the judgments.

Multimodal data and even classification models have uncertainty, which may be categorized as data uncertainty and model uncertainty. Noise and uneven data volume are uncertainty in voice data for sentiment recognition, whereas lighting, contrast, and shooting angle are uncertainties in picture data. There is no possibility to simulate a totally correct recognition model with limited parameters due to model uncertainty. Murat Sensoy et al. introduced evidential deep learning for measuring uncertainty (Murat Sensoy,2018) because uncertainty is pervasive in data and models but difficult to describe and measure correctly. As the Softmax function removes the trustworthiness of each modality, the Softmax of the final output layer must be replaced with alternative activation functions whose outputs are not negative. In this article, the Relu activation function is used, and the Relu function’s output serves as the evidence layer. $e = [e_1, e_2, \dots, e_k]$, and $k = 4$, which reflects the four types of emotion discussed in this paper? The related categories in the evidence layer no longer indicate the likelihood of recognition, but rather the quality of evidence, which may quantitatively describe the classification model’s dependability and will be utilized for future probability distribution and uncertainty computation. The formula for its computation is as follows:

$$u = \frac{c}{\sum_{i=1}^k (e_i + 1) + c} \tag{1}$$

where: u is the uncertainty, c is the defined constant, and $\sum_{i=1}^k e_i$ is equal to the sum of the all-evidence mass in the evidence layer.

Under constant circumstances, the greater the sum of the quality of the evidence in the output layer, the lower the value that shows the uncertainty of the modal choice. As uncertainty is the inverse of the modality's credibility, this also signifies the modality's greater credibility. Following analysis, this phrase may effectively depict the relative size of uncertainty. The output evidence layer then models the choices of the two modalities individually $e^{au} = [e_1^{au}, e_2^{au}, \dots, e_k^{au}]$ for the audio emotion recognition model and evidence output layer $e^{ima} = [e_1^{ima}, e_2^{ima}, \dots, e_k^{ima}]$ Given the image recognition model, then use Eq. 5 to independently construct the two evidence vectors, and the credibility of the audio recognition decision is $1 - u^{au}$, and the credibility of the image recognition decision is $1 - u^{ima}$. Finally, we calculate the probability distribution of the decision:

$$p_t = \frac{e_t + 1}{\sum_{i=1}^k (e_i + 1) + c} \quad (2)$$

where: p_t is the predicted probability value of the decision for the type t of outcome, and $\sum_{t=1}^k p_t + u = 1$ is derived from Eq. 2.

2.3 Decision Fusion Based on D-S Evidence Theory

The identification framework is the fundamental concept of this decision information fusion theory, which contains independent belief quality distributions in which all framework elements are mutually exclusive and ensures that the framework contains all possible outcomes and a finite number of outcomes, as shown in Eq. 3.

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_N\} \quad (3)$$

where Θ is the identification frame, θ is each individual element of the frame.

The four independent components in this article correlate to the four classed emotions "happy," "sad," "terrified," and "neutral." The four emotions are autonomous and mutually exclusive in the categorization trials, and each emotion is referred to as an element of the emotion identification framework Θ in this paper. Then, another fundamental term in evidence theory, the power set, is likewise a set consisting of all subsets of the recognition framework, with the following expression:

$$2^\Theta = \left\{ \Phi, \{\theta_h\}, \{\theta_s\}, \{\theta_f\}, \{\theta_n\}, \{\theta_h, \theta_s\}, \{\theta_h, \theta_f\}, \dots, \{\theta_h, \theta_s, \theta_f, \theta_n\} \right\} \quad (4)$$

where: $\theta_h, \theta_s, \theta_f, \theta_n$ corresponds to the independently distributed probabilities of each one from the four emotions, $\{\theta_h, \theta_s\}$ represents the fuzzy probability that the emotion is "happy" or "sad", $\{\theta_h, \theta_s, \theta_f, \theta_n\}$ is the full set of the recognition framework, and represents the fuzzy probability that all four emotions are possible:

$$p(A) = \begin{cases} \frac{\sum_{A_i \cap B_j} p^{au}(A_i) p^{ima}(B_j)}{1-K}, & A \neq \Phi \\ 0, & A = \Phi \end{cases} \quad (5)$$

$$K = \sum_{A_i \cap B_j = \Phi} p^{au}(A_i)p^{ima}(B_j) < 1 \tag{6}$$

where: K is the conflict factor, $1/(1 - K)$ is the regularization factor, and the destination equation lets the sum of expressions be 1.

Through the steps in the above, the audio recognition decision distribution $p^{au}[p_1^{au}, p_2^{au} \dots p_k^{au}, u^{au}]$, and the image recognition decision distribution $p^{ima}[p_1^{ima}, p_2^{ima} \dots p_k^{ima}, u^{ima}]$ are now derived. $p_1, p_2 \dots p_k$ where is the reasonable probability distribution for each kind of emotion and represents the uncertainty. When a modality is unsure about the ultimate choice, it may be seen as a fuzzy probability, both among the uncertainty probabilities, which may be the probability of each feeling distribution, and the whole set of probabilities Θ in the power set, which corresponds $\{\theta_h, \theta_s, \theta_f, \theta_n\}$ in this paper.

If the conflict coefficient K is equal to 1, there is no way to calculate it. When the K value is too large and the conflict between propositions is too obvious, it is feasible to construct paradoxes by fusing illogical choice information fusion outcomes.

Figure 3 depicts how we then combine the multimodal choice information using D-S evidence theory.

The total of the empty regions in Fig. 3a shows the K -conflict coefficient, which is the distribution of mutually exclusive modal choices. For instance, when the audio recognition model identifies “happy” emotion and the image recognition model identifies “sad” emotion, the two modalities result in contradicting conclusions, and when the audio recognition model identifies “sad” emotion, the two modalities result in conflicting decisions. When the audio recognition model identifies a result as “sad” and the picture recognition model identifies the outcome as “uncertain,” the fusion judgment “sad” is not contradictory and may be deemed logical reasoning. In accordance with the evidence theory, conflicting information is deleted and non-conflicting information is re-normalized in order to get a fusion judgment. If we conclude directly in this manner, we will find that four categories of emotions were initially identified, but the classification result also includes a category of “uncertainty.” Therefore, in the fusion process,

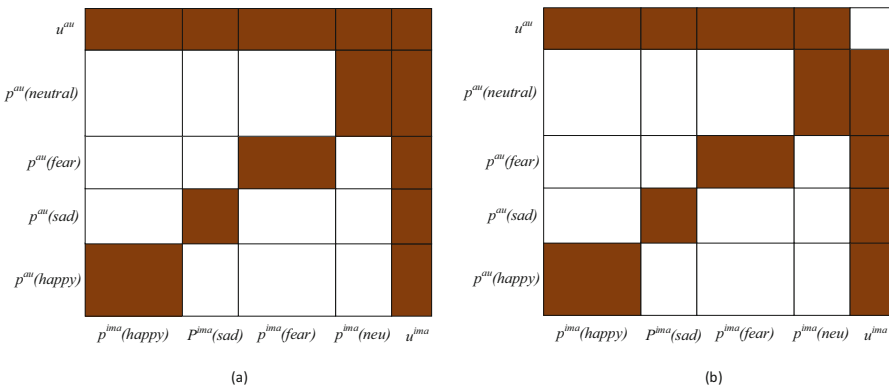


Fig. 3. Decision information fusion floor plan(a)/(b)

this paper will also include the case of uncertainty in both modalities into the conflict coefficient, as depicted in Fig. 3 (b), and the new method for calculating the conflict coefficient is shown in Eq. 7:

$$K' = \sum_{A_i \cap B_j = \Phi} p^{au}(A_i)p^{ima}(B_j) + u^{au}u^{ima} < 1 \quad (7)$$

3 Experiments

This work discusses the enhancement of the modal output layer by substituting the Softmax activation function with the Relu activation function. The findings of the output layer are employed as the evidence layer for subsequent credible multimodal fusion. c value is very important, and this chapter adjusts the size of c via experiments to acquire varied recognition effects on MELD datasets (Soujanya Poria,2019), and then concludes. The experimental outcomes are shown in Table 1.

From Table 1, it can be deduced that the accuracy does not grow or decrease constantly as the constant c changes. Instead, there is a peak at $c = 6$, with a model accuracy of 81.25 percent. The magnitude of the uncertainty is affected by the value of the constant c . If the value of c is too tiny, the uncertainty is negligible, and the extracted uncertainty plays no part in the fusion procedure. If the value of c is big and the model uncertainty is high, it is impossible to make a credible judgment, and the whole model effect will be deteriorated, resulting in a loss in classification accuracy.

From Table 2, it can be seen that the accuracy obtained by decision fusion using the Softmax activation function directly with the mean is even less than that obtained by unimodality. This is because, after using the Softmax activation function, the belief quality information of the modality is discarded and the recognition rate of almost all categories is too high, resulting in “overconfidence” of unimodality. If the modality is misclassified, it will contaminate the decision of the unimodality. In contrast, the Relu activation function, despite its lower correct rate improvement, is more rational and

Table 1. Different c values correspond to the fusion accuracy

Condition	Accuracy on MELD datasets
$c = 1$	73.25
$c = 2$	73.75
$c = 3$	74.00
$c = 4$	76.25
$c = 5$	78.00
$c = 6$	81.25
$c = 7$	79.25
$c = 8$	79.00

Table 2. Recognition accuracy under different fusion algorithms

Fusion method	Recognition accuracy
Softmax + Average method	75.00
Relu + Average method	79.75
Softmax + Cascade + FC	78.50
Relu + Cascade + FC	79.50
Method in this paper	81.25

capable of producing consistent improvement. After removing the uncertainty of the modalities, we apply evidence theory to further fuse the decision information and get an 81.25% right identification rate, which also demonstrates the applicability of the fusion technique introduced in this chapter for multimodal emotion detection.

4 Conclusion

This research offered audio and image-based algorithms for sentiment recognition. Next, the output of the classification model is re-modeled by substituting the activation function, from which the uncertainty of the modal correspondence choice is derived, and the decision information fusion process of the multi-modal re-modeling is adopted by the D-S evidence theory. In order to address the possibility of paradoxes occurring in the actual application of evidence theory or the impossibility of doing the calculation, this chapter proposes a new method for assigning belief quality as part of re-modeling the evidence layer. In the experimental process, the optimal value of uncertainty extraction constant c is first determined through controlled experiments. Then, the effect of the multimodal fusion algorithm proposed in this chapter is compared with the unimodal recognition algorithm, and the results of various fusion methods are compared. Finally, a recognition accuracy of 81.25 percent is achieved on the MELD dataset, indicating that the fusion algorithm in this chapter is feasible and effective.

References

1. Abdullah S M S A, Ameen S Y A, Sadeeq M A M, et al. Multimodal emotion recognition using deep learning[J]. Journal of Applied Science and Technology Trends, 2021, 2(02): 52-58.
2. Chen M, He, Yang J, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10):1440-1444.
3. D’Mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems [J]. ACM Computing Surveys, 2015,47(3) :1-36.
4. Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems, pp. 3179-3189, 2018.
5. Noda K, Arie H, Suga Y, et al. Multimodal integration learning of robot behavior using deep neural networks [J]. Robotics and Autonomous Systems ,2014,62 (6) :721-736.

6. Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
7. Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines[J]. *Journal of Machine Learning Research*, 2014, 15(8) :2949-2980.
8. Ortega J D S, Senoussaoui M, Granger E, et al. Multimodal fusion with deep neural networks for audio-video emotion recognition[J]. arXiv preprint [arXiv:1907.03196](https://arxiv.org/abs/1907.03196), 2019.
9. Shen G, Lai R, Chen R, et al. WISE: Word-Level Interaction-Based Multimodal Fusion for Speech Emotion Recognition[C]//Interspeech. 2020: 369–373.
10. Nemati S, Rohani R, Basiri M E, et al. A hybrid latent space data fusion method for multimodal emotion recognition[J]. *IEEE Access*, 2019, 7: 172948-172964.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

