Intelligent Computing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Tracking Emotions Using an Evolutionary Model of Mental State Transitions: Introducing a New Paradigm

Fu-Ji Ren[1], Yang-Yang Zhou[2], Jia-Wen Deng[1], Kazuyuki Matsumoto[2], Duo Feng[2], Tian-Hao She[2], Zi-Yun Jiao[1], Zheng Liu[2], Tai-Hao Li[3], Satoshi Nakagawa[4], and Xin Kang[2*]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. [2]Department of Computer Science, Tokushima University, Tokushima, Japan. [3]Research Center for Multi-Modal Intelligence, Research Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou, China. [4]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan.

*Address correspondence to: kang-xin@is.tokushima-u.ac.jp

Citation: Ren FJ, Zhou YY, Deng JW, Matsumoto K, Feng D, She TH, Jiao ZY, Liu Z, Li TH, Nakagawa S, et al. Tracking Emotions Using an Evolutionary Model of Mental State Transitions: Introducing a New Paradigm. *Intell. Comput.* 2024;3:Article 0075. https://doi.org/10.34133/icomputing.0075

Submitted 29 May 2023
Accepted 29 November 2023
Published 8 April 2024

Owing to rapid advancements in artificial intelligence, the role of emotion recognition has become paramount in human–computer interaction. Traditional approaches often reduce this intricate task to a mere classification problem by relying heavily on perceptual pattern-recognition techniques. However, this simplification overlooks the dynamic and multifaceted nature of human emotions. According to theories in emotion psychology, existing pattern recognition methods primarily capture external emotional expressions—termed "external emotional energy" (EEE)—rather than the nuanced underlying emotions. To address this gap, we introduce the evolutionary mental state transition model (EMSTM). In the initial phase, EMSTM employs standard pattern-recognition algorithms to extract EEE from multi-modal human expressions. Subsequently, it leverages a mental state transition network to model the dynamic transitions between emotional states, thereby predicting real-time emotions with higher fidelity. We validated the efficacy of EMSTM through experiments on 2 multi-label emotion datasets: CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) and Ren Chinese Emotion Corpus (Ren-CECps). The results indicate a marked improvement over conventional methods. By synergistically combining principles from psychology with computational techniques, EMSTM offers a holistic and accurate framework for real-time emotion tracking, aligning closely with the dynamic mental processes that govern human emotions.

Introduction

Emotion, as a complex and advanced form of human intelligence, has garnered attention in the field of human–computer interaction with the development of artificial intelligence technology. Research on emotion has become crucial for enabling machines to understand and possess emotional capabilities, such as emotion recognition, expression, and generation, in various information fields [1–3].

Emotion recognition is the process by which a machine identifies human emotions, and it serves as the foundation for giving machines emotional capabilities [4]. Emotion recognition has been studied for decades, and the field has seen substantial theoretical advancements. Automatic emotion recognition systems have applications in various domains of daily life, such as public opinion monitoring [5], marketing communications [6], and mental health monitoring [7]. In the context of human–computer interaction [8], emotion recognition technology plays a crucial role in enhancing user experience. Emotion tracking involves the continuous detection of emotions over a period of time. Real-time emotion tracking, particularly in online education, has shown the potential to improve students' learning motivation and effectiveness [9].

Current emotion recognition technology predominantly relies on pattern-recognition methods that involve the extraction of emotional features from external information sources [10]. These sources include observable physical behaviors [11,12] and monitorable physiological information [13,14]. By extracting and combining features from different modalities, a coherent representation is created that enables further emotion inference. In several cases, this inference process has involved multi-category classification tasks aimed at predicting the presence of specific emotions [15,16].

However, human emotion is a dynamic and intricate phenomenon [17]. The emotions recognized by most existing pattern-recognition methods only scratch the surface of the complex nature of human emotion. According to Freud's psychoanalytic theory [18], psychic energy is a psychological force that drives individuals to exhibit appropriate behaviors and experience a range of emotions based on their internal subjectivity. We propose that observable behaviors are expressions of this psychic

emotional energy, which we refer to as external emotional energy (EEE). EEE is generated through the process of emotional perception and directly influences the dynamic transitions between mental states. That is, external behaviors serve as observable manifestations [19] that reflect EEE and provide realistic feedback on its impact on mental states. In this context, we define mental state as the state of each emotion and transition as the process by which mental states change under the influence of EEE. Hence, existing pattern recognition methods primarily recognize EEE rather than capturing the essence of real emotions.

In complex and advanced human intelligence, emotions involve both cognitive and physical aspects [20]. Simulating a mental state by relying solely on external physical information is challenging. The associations between modalities and the integration of context and of historical emotions must be explored to track internal emotions. This requires an intelligent emotion recognition system that encompasses integrated processes, including pattern recognition, historical emotion tracking, and mental state transition. We define the updating of recognized emotions based on historical emotions as an evolutionary mental state transition process. This study aims to construct an evolutionary emotion-tracking system and validate its performance.

In this study, we propose a new paradigm for emotion tracking called the evolutionary mental state transition model (EMSTM). EMSTM aims to simulate the process of tracking a person's mental state transition using observable behaviors and historical emotions. It incorporates pattern recognition methods and introduces the concept of a mental state transition network (MSTN) from our previous study [21]. A conceptual diagram of EMSTM is shown in Fig. 1, where the pattern recognition method calculates the EEE embodied in multi-modal expressions. The MSTN represents the manner in which EEE influences mental state transition.

The contributions of this study are as follows:

1. We construct the EMSTM system using artificial intelligence techniques to simulate the association between multi-modal emotion expressions and mental state transition.
2. We propose a new pattern recognition method that establishes associations across various modalities of emotional expressions, enabling the calculation of EEE driving multi-modal behavior.
3. Based on the concept of EEE, we introduce MSTN to simulate the dynamic emotional transition process.

We conducted experiments on 2 multi-label emotion classification datasets to validate the effectiveness of the proposed paradigm. The remainder of this paper is organized as follows: Materials and Methods provides a detailed description of our proposed model by combining insights from related studies with the methodologies employed in EMSTM. Results and Discussion presents an analysis of the performance of the system. Finally, Conclusion concludes the paper and outlines future research directions.

Materials and Methods

Methodological framework and theoretical foundations

Emotion recognition research has evolved from single-modal to multi-modal analysis, acknowledging the complexity of accurately capturing human emotions [22]. Multi-modal approaches, which incorporate information from various modalities, address the limitations of single-modality methods by providing a richer context for emotion prediction. This development has been crucial in overcoming the biases and misinterpretations that arise from relying on a single source of information.

Multi-modal fusion methods have emerged as key strategies in this field, with 2 predominant approaches: feature-level fusion

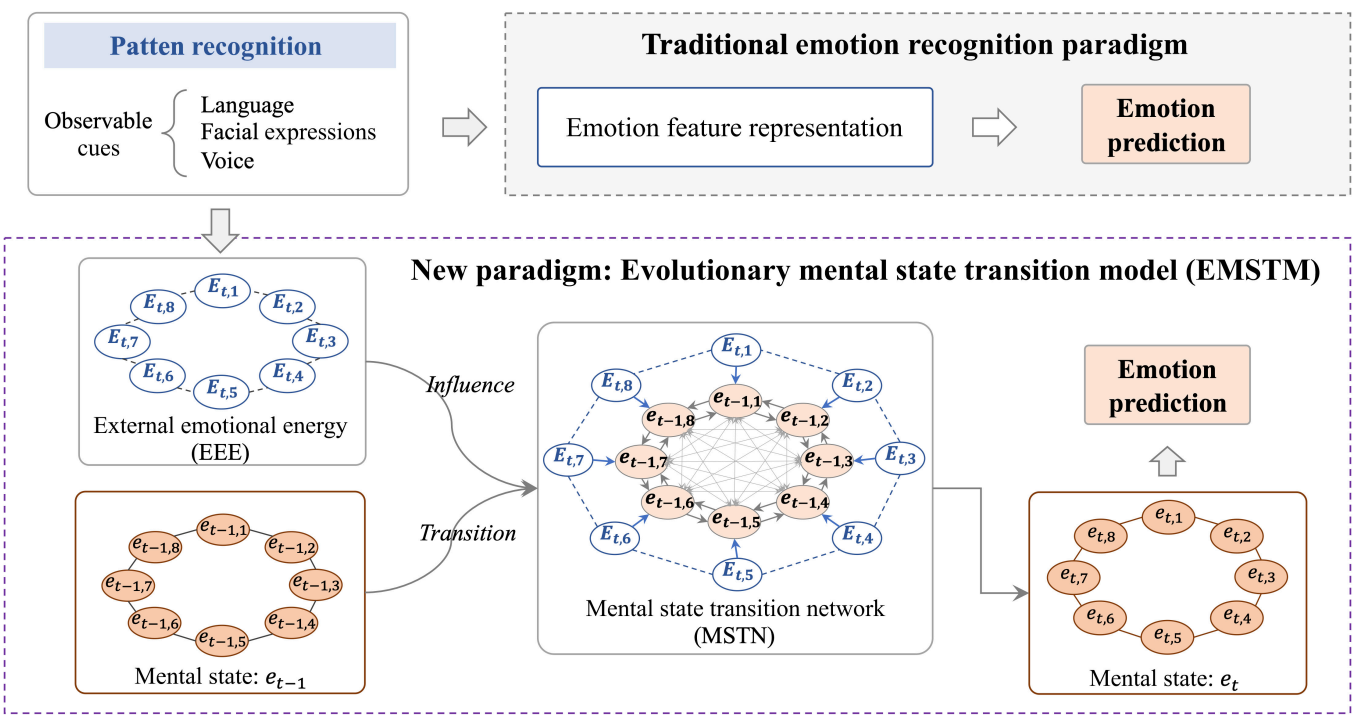

Fig. 1. Conceptual diagram of the proposed evolutionary mental state transition model (EMSTM).

and decision-level fusion [23]. Feature-level fusion involves combining features from each modality into a unified representation to enhance the comprehensiveness of emotional analysis [24,25]. In contrast, decision-level fusion processes each modality independently and combines their outputs, emphasizing the unique contributions of each modality to the overall emotion prediction [26–28].

Complementing these strategies is the concept of dynamic emotional transition, which focuses on the patterns and regularities of emotional fluctuations over time [17]. This approach extends beyond static emotion analysis by considering how emotions evolve and influence each other, particularly in the context of multiparty interactions [29–32]. Recent advances in deep learning have further enriched this domain with models that leverage hierarchical networks and RNN-based systems for the context-level modeling of dynamic emotions [16,33–44].

Psychoanalytic theory suggests that emotional transitions are not merely external behavioral changes but represent deeper internal processes within an individual's mental state [21,45–51]. This perspective is crucial for understanding the dynamic nature of emotions and their impact on mental states.

In response to these evolving dynamics in emotion research, we propose the EMSTM, which is a comprehensive framework designed to model the dynamic interplay between observable multi-modal behaviors and mental states. EMSTM integrates several key components. It starts with multi-modal pattern recognition, where we analyze and encode features from various modalities including language, vision, and acoustics. This is followed by cross-modality feature fusion, which employs advanced techniques to synthesize information across these modalities. The process then moves to the calculation of EEE, which quantifies the emotional energy from multi-modal data, and is essential for the subsequent steps. The core of the model is the MSTN, which simulates the manner in which EEE influences mental states over time and models the dynamic transitions of emotions. Collectively, these components work cohesively to provide a nuanced understanding of emotional transitions based on multi-modal data analysis.

Multi-modal pattern recognition

In dynamic psychology, the concepts of dynamic rules and energy have been introduced in personality research [45]. The human mental system is considered to be a dynamic system that requires energy, similar to other physical dynamic systems. This energy is referred to as EEE. In the external environment, EEE is manifested through observable psychological behaviors, particularly in the form of multi-modal emotional expressions. EEE influences the mental state, and its generation and transfer play a role in emotional transitions. Consequently, we propose the use of EEE as an intermediary to simulate the relationship between multi-modal emotional expressions and the internal subjectivity of human emotional inference.

This study focused on the analysis of multi-modal expressions, including language, vision, and acoustic modalities. The generation and transfer of EEE are dynamic processes that are influenced by 2 main factors.

1. Externally triggered: EEE is produced when individuals perceive information and emotions from the outside world, such as when engaging in interactions or receiving external stimuli.

2. Internally triggered: EEE arises from self-emotional expressions, such as self-monologue or changes in one's thoughts.

Whether triggered externally or internally, the generated EEE is reflected in observable psychological behaviors, primarily through multi-modal expressions. Fluctuations in EEE impact emotional transitions, and emotions are simultaneously externalized through subsequent behaviors.

To simplify the calculation of EEE, we modeled the recursive process as follows: the emotional transition at the current time is affected only by the EEE at the previous moment, and the EEE at the current time affects only the subsequent emotional transition.

In this study, we considered 3 modalities: language (l), vision (v), and acoustics (a). If we conceptualize a sentence as a moment, the task can be defined as follows: Given the external information $x'_{m,t}$ of modality m at time t, we obtain its underlying feature representation $x_{m,t}$ as follows:

$$x_{m,t} = g\left(x'_{m,t}\right), \quad m \in [l,v,a]. \tag{1}$$

The goal is to calculate EEE E_t from $x_{m,t}$, which can be described as follows:

$$E_t = f\left(x_{m,t}\right). \tag{2}$$

Mono-modality feature encoding

Various methods are available to represent these modalities, most of which are compatible with our proposed model. However, in this study, we did not focus on discussing which method was superior. At each time step t, we utilized the feature extraction methods (referred to as $g()$ in this study) described in [52] to obtain the underlying features of the language, vision, and acoustic modalities $x_{l,t}$, $x_{v,t}$, and $x_{a,t}$, respectively, as shown in Fig. 2.

Padding operations were employed to ensure that the sequence lengths of each modality were consistent. For example, we padded the token length of the current sequence to a fixed sequence length L_m. We employed 3 additional fully connected networks to standardize the feature dimensions of the 3 modalities. Consequently, the shape of the input features $x_{m,t} \in \mathbb{R}^{L_m \times D_m}$, where L_m represents the sequence length and D_m denotes the feature dimension of each modality.

Based on our formulation, the calculation of EEE relies on sentence-level representations. However, the underlying features we obtained were character-level features that contained positional information. We did not employ recursive or positional embedding methods on character-level features when fusing modalities. This differentiates our approach from some of the baseline models used in the experiment, such as Mu-Net.

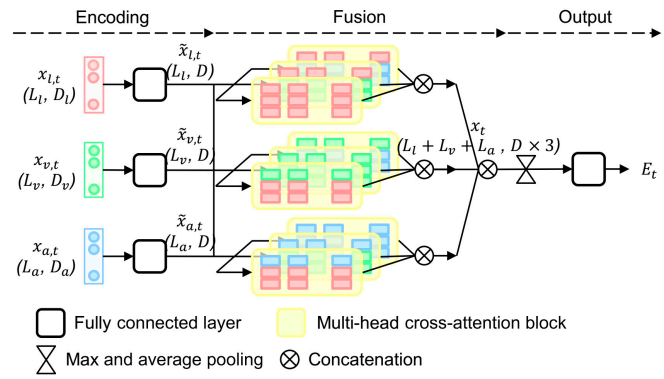

Fig. 2. Overall framework of the multi-modal external emotional energy (EEE) calculation model.

To encode the underlying features, we applied a simple linear projection to unify the feature dimension as D. The encoding features $\tilde{x}_{m,t}$ are expressed as

$$\tilde{x}_{m,t} = x_{m,t} W_u, \tag{3}$$

where W_u represents the weight of the fully connected layer used for dimension unification. The shape of the encoding features $\tilde{x}_{m,t}$ is (L_m, D).

Cross-modality feature fusion

The multi-head self-attention transformer was initially introduced by Google Brain [53]. It utilizes a query (Q) to calculate the similarity between the query and each key (K) and then computes the weighted sum of all values (V) based on this similarity. This mechanism is commonly employed to emphasize important information within the encoding features. Motivated by this mechanism, we employed 9 multi-head cross-attention blocks to determine the importance of the encoding features across modalities.

As shown in Fig. 3, we modified the original block to facilitate the deployment of the model on robots or mobile devices. Considering that fully connected networks are used for linearly projecting input features during the encoding process, we eliminate the linear projection of QKV employed in previous transformer-based studies. For instance, we utilize the base-modality feature $\tilde{x}_{v,t}$ as Q in this architecture, and the cross-modality feature $\tilde{x}_{a,t}$ serves as both K and V. The cross-attention feature between v and a is described as follows:

$$x_{vaa,t} = \mathrm{Softmax}\left(\frac{\tilde{x}_{v,t}\tilde{x}_{a,t}^{\mathrm{T}}}{\sqrt{D}}\right)\tilde{x}_{a,t}, \tag{4}$$

where $\tilde{x}_{v,t}, x_{vaa,t} \in \mathbb{R}^{L_v \times D}$ and $\tilde{x}_{a,t} \in \mathbb{R}^{L_a \times D}$. The base-modality encoding feature is represented in the feature space of the cross-modality.

To maintain consistency with the proposed MSTN, we employed a skip connection to concatenate features from different representation spaces within the original block. In contrast to previous studies, we concatenated the base-modality features with the cross-attention features along the feature-dimension axis. We then utilized a fully connected layer to restore the feature dimension to D and ensure the continuity of the blocks. In the example mentioned earlier, the fused feature of v and a is expressed as

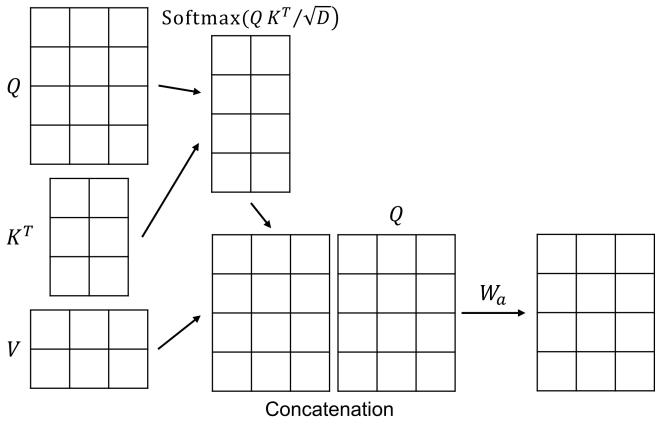

Fig. 3. Architecture of the multi-head cross-attention block.

$$x_{va,t} = \mathrm{Concat}\left(\tilde{x}_{v,t}, x_{vaa,t}\right) W_a, \tag{5}$$

where $x_{vaa,t}$ represents the cross-attention feature between v and a, W_a denotes the weight of the fully connected layer used for dimension adjustment, and $x_{va,t} \in \mathbb{R}^{L_v \times D}$.

A similar fully connected structure was utilized in the position-wise feedforward network, followed by another skip connection, as employed in previous transformer studies. Using a small amount of data, we verified that removing the position-wise feed-forward network did not affect the convergence of the loss value in this model.

As shown in Fig. 2, we sequentially concatenated the 9 fused features along both the feature-dimension and sequence-length axes. The base modalities and cross-modalities of the 9 features are l & a, l & v, l & a, v & l, v & v, v & a, a & l, a & v, and a & a. The shape of the hidden features x_t after concatenation is $(L_l + L_v + L_a, D \times 3)$.

External emotional energy calculation

Pooling is a commonly used technique in deep learning for dimensionality reduction through downsampling. In our approach, we utilized maximum pooling to retain the relative positional relationships of important features and average pooling to capture background information. After dimensionality reduction, the hidden features were transformed into EEE through an unbiased, fully connected layer as follows:

$$E_t = x_t W_e, \tag{6}$$

where E_t represents the EEE, and its shape is (n), where n is the number of emotion categories. EEE is used for emotion calculation in the MSTN. In this study, we employed an unbiased, fully connected layer to ensure uniform performance across all emotion categories. Although biased classifiers have been shown to excel in incremental learning scenarios, as discussed in Wu et al. [54], our methodology was designed to achieve balanced classification without giving undue preference to newer classes.

Mental state transition network

In psychoanalytic theory, emotion recognized through pattern recognition methods is more appropriately termed EEE. EEE influences the mental state rather than directly representing real human emotion.

We introduced the concept of the MSTN to simulate the influence of EEE on the mental state and predict the dynamics of emotional transitions. MSTN aims to capture the flow of EEE, leading to changes in historical emotions and, consequently, an updated mental state.

As illustrated in Fig. 4, MSTN models the transfer and redistribution of EEE within the mental state, resulting in emotional transitions over time. The energy transfer from EEE to the mental state leads to changes in emotions, which can be simulated using MSTN to obtain the updated mental state.

To model the mental state transition, we assume that the mental state at time t, denoted by e_t, consists of multiple discrete emotion categories:

$$e_t = \left(e_{t,1}, e_{t,2}, \cdots, e_{t,n}\right), \tag{7}$$

where t represents the time step, and n is the number of emotion categories. Each element $e_{t,i}$ represents the active intensity of emotion e_i at time t, scaled within a specific range, such as $[0, 1]$.

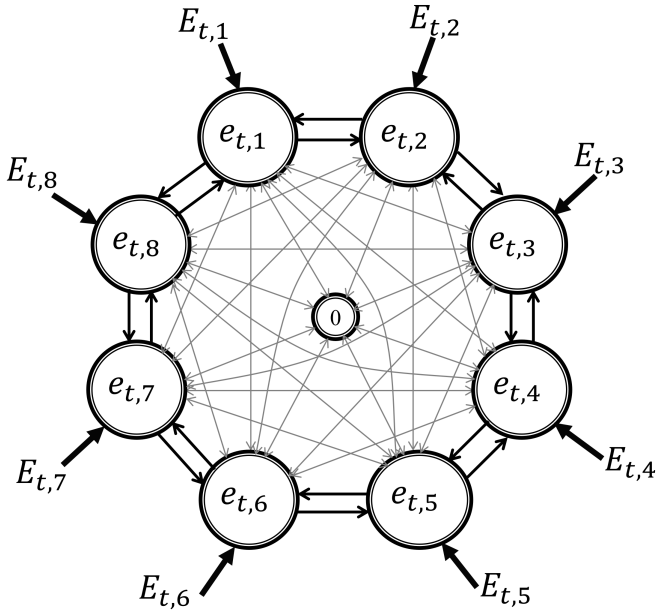

Fig. 4. Conceptual diagram of the mental state transition network (MSTN), the number of emotion categories $n = 8$.

Similarly, EEE can be represented as discrete categories as follows:

$$E_t = \left(E_{t,1}, E_{t,2}, \cdots, E_{t,n}\right). \tag{8}$$

As shown in Fig. 4, the energy E_t influences the mental state e_t. The central circle represents the neutral state, whereas the surrounding circles represent other emotional states. The weighted arrows indicate the probabilities of emotions transitioning from one state to another. This is the underlying principle of MSTN: transferring energy from the EEE to the mental state, resulting in emotional transitions.

In this study, we extend our previous work [30], in which we collaborated with psychologists to conduct a psychological questionnaire survey involving 200 participants and established a preliminary transition probability matrix for MSTN. The construction process of MSTN is elaborated in the preceding sections. Unlike the model used in our previous study, which treated emotion as context-independent discrete jumps, we now model emotion as a dynamic mental state influenced by changes in EEE. This allows the tracking of emotional transitions over time, making MSTN more aligned with real-world emotional dynamics. The key improvement in this study is the introduction of EMSTM, which refines the emotion transition links within MSTN. Our approach aims to address unanswered scientific questions related to evolutionary emotional cognition and individual emotional differences.

It can be verified that, under certain conditions, the mental state can transition from one state to another. Although the transition probabilities among the states are not equal, we believe that, in the absence of external factors, there is an expected probability. Therefore, an MSTN can be constructed based on the analysis of extensive data and personality information.

Human dynamic emotions exhibit a nondeterministic pattern, and a Markov state chain is suitable for modeling this pattern. Traditional Markov chains assume that the current mental state depends solely on the previous state. However, emotional transition is a more complex process involving energy transfer and redistribution among emotions and is influenced by EEE. The updated mental state becomes the current mental state. To account for this, we assume that the current mental state depends not only on the previous state but also on the final mental state and the EEE. The emotional transition process in MSTN can be expressed as follows:

$$P\left(e_t | e_{t-1}, e_{t-2}, \cdots, E_t, E_{t-1}, E_{t-2}, \cdots\right) = P\left(e_t | \tilde{e}_{t-1}, E_t\right), \tag{9}$$

where e_t represents the updated mental state affected by EEE E_t in the previous mental state after transitioning $\tilde{e}_{t-1}$.

The emotions expressed by individuals at a given moment can be determined based on the active intensity of each emotion in the mental state. If the intensity of a particular emotion exceeds a threshold, we can infer that this emotion is activated in the mental state and is, therefore, included in the current expression. Compared to a mono-emotion classification system, this multi-emotion recognition mechanism provides a more comprehensive inference of the mental state.

The human mental system can be likened to a dynamic system in which real emotions depend on historical emotions and EEE, described as follows:

$$e_t = \mathrm{Concat}\left(\tilde{e}_{t-1}, E_t\right) W_c, \tag{10}$$

where W_c denotes the weights of the fully connected layer classifier.

Considering personality, the extent to which the mental state is affected by EEE varies across situations. Therefore, the EEE and final mental state involved in the above formula are weight-adaptive.

There are different types of associations between emotions, such as the positive association between joy and love and the negative association between love and hate. These emotional associations influence the transfer of emotions within the mental state and occur with certain probabilities. To account for this factor, we define an emotional transition matrix T. In the case of n emotion categories, the matrix is denoted by $T \in \mathbb{R}^{n \times n \times n}$, which describes the transition probabilities between emotions. Each element T_{ijk} denotes the probability of energy transfer from historical emotion e_j to current emotion e_k under the influence of emotion e_i represented by the energy.

The mental state at the current moment is transitioned from the previous mental state through matrix T. At this moment, the EEE activates the previous mental state based on its weight.

$$\tilde{e}_{t-1} = \mathrm{Norm}\left(E_t \times \left(e_{t-1} \times T\right)\right), \tag{11}$$

where Norm normalizes the transition and activation processes to prevent the gradient from vanishing or exploding.

By calculating EEE E_t using a multi-modal pattern recognition method and employing the emotional transition matrix T to compute the updated mental state $\tilde{e}_{t-1}$ based on the previous real emotion e_{t-1}, we can determine the current real emotion e_t. This approach presents an emotion-tracking paradigm that aligns closely with psychological cognition, going beyond treating observable behaviors as mere expressions of real emotions.

Evolutionary mental state transition model

In this study, we develop the MSTN described above and simulate emotional transitions by tracking the fluxion of EEE. EEE is derived from observable multi-modal behaviors. To establish

a clear connection between the mental state and multi-modal behaviors, we employed deep learning technology to construct an EMSTM that simulates the interaction between the mental state and the external environment.

The overall framework of EMSTM is illustrated in Fig. 5. EMSTM predicts dynamic changes in the mental state over time by simulating the influence of EEE. It first calculates the EEE from multi-modal data and then predicts the mental state using the MSTN. The detailed process is as follows:

1. Calculation of EEE: This step involves calculating the EEE based on the emotional interactions expressed in multi-modal data. Please refer to the "Multi-modal pattern recognition" section for a detailed explanation.

2. Inference of updated mental state: In this step, the emotional transition process is simulated using the proposed MSTN to predict the updated mental state in real time. Please refer to the "Mental state transition network" section for further details.

EMSTM can simulate the dynamic emotional transition process when provided with a sequence of multi-modal emotion expression data. The input to EMSTM consists of multi-modal information, including language, vision, and acoustic features, arranged in chronological order. EMSTM first obtains sequential multi-modal embeddings using a multi-modal encoder. It then calculates the EEE from these embeddings at each time step. This EEE, serving as emotional energy, interacts with the MSTN, resulting in the updating of the mental state. By comparing the predicted mental state with the ground truth in the time series, EMSTM can be optimized by minimizing the error.

Considering that the parameters and computational effort of the model primarily lie in feature fusion, we calculated the EEE independently, as shown in Fig. 5. This approach accelerates model computation in parallel. Additionally, we updated the current mental state based on the EEE and previous mental state instead of directly referencing the long-term past state. Unlike traditional methods of stacking recurrent neural networks (RNNs), we modeled the temporal sequence in the mental state using fewer parameters. The experimental results demonstrate the effectiveness of the proposed model.

Results and Discussion

Results on CMU-MOSEI

Dataset and evaluation metrics

We conducted experiments on the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset to validate the effectiveness of our model for emotional multilabel classification. The dataset consisted of 1,000 individual monologues with 23,453 annotated data points covering 6 emotions: happiness, sadness, anger, disgust, surprise, and fear. The raw data comprised approximately 66 h of monologue videos. The multi-modal feature vectors included 35 facial action units recorded by Facet, acoustic features extracted using Cooperative Voice Analysis Repository (COVAREP), and language features obtained using the GloVe embedding method. It is important to note that the original publication of the dataset ([52]) does not specify whether a cross-subject paradigm was employed for data splitting. Owing to the anonymized nature of the dataset, we could not confirm this. However, we strictly adhered to the original authors' guidelines for data splitting, ensuring that the train/test splits were consistent with those used for the baseline methods. For additional details on the multi-modal features and data splitting methodology of the dataset, please refer to the work of Zadeh et al. [52].

Figure 6 illustrates the distribution of the emotion labels in the CMU-MOSEI dataset. The horizontal and vertical axes represent the 7 emotions, including neutral. The cells on the diagonal indicate the total number of occurrences of each emotion. Happiness stands out with a larger quantity compared to the other emotions. The other cells represent the co-occurrence counts of 2 different emotions. Co-occurrences involving more than 2 emotions are not represented in the figure.

For evaluation, we utilized standard accuracy and weighted F1 scores calculated using the sklearn tool [55]. The F1 score provides a balanced measure of precision and recall and is particularly useful for imbalanced datasets with categories that have a small amount of data. In addition, we computed the weighted accuracy (*wa*) to evaluate the overall performance of

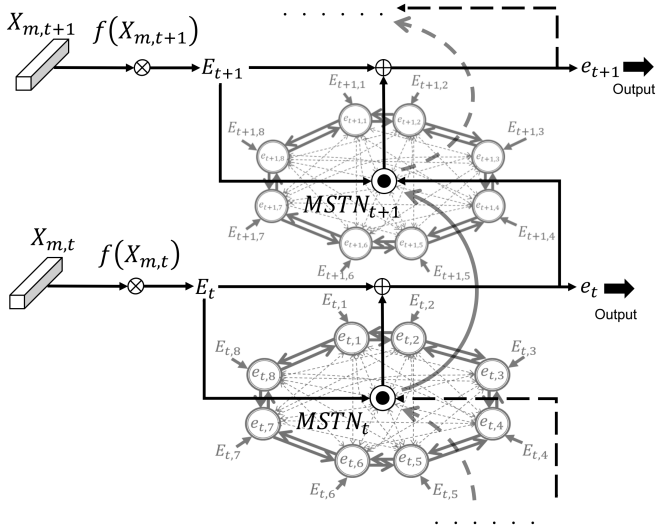

Fig. 5. Overall framework of the EMSTM.

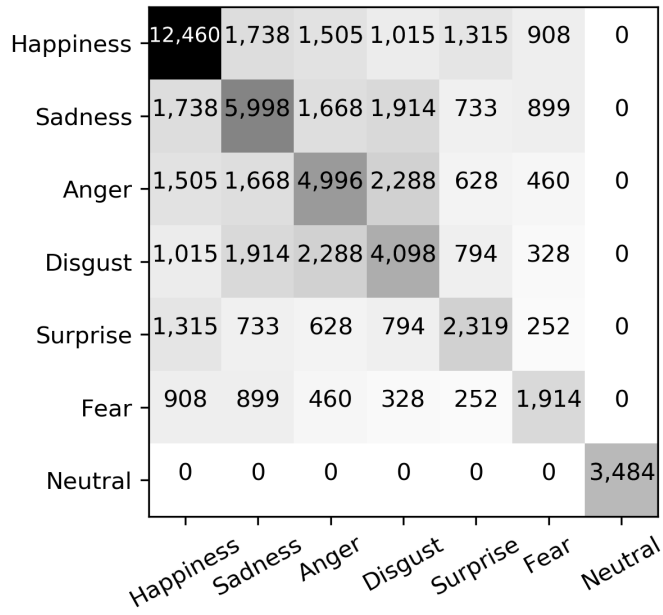

Fig. 6. Distribution of emotion labels in the CMU-MOSEI dataset.

the models. *wa* is calculated based on the amount of data associated with different emotions.

$$wa = \frac{\sum_{i=1}^{n} s_i a_i}{\sum_{i=1}^{n} s_i}, \quad (12)$$

where n denotes the total number of emotions, s_i denotes the amount of data containing emotion i, and a_i denotes the accuracy of emotion i. Note that the evaluation metrics we used do not consider the neutral emotion. The weighted accuracy figures for Deep Emotional Arousal Network (DEAN), Feature-Disentangled Multimodal Emotion Recognition (FDMER), and Modality Co-Reinforcement (MCR) do include the neutral emotion, as reported in their respective papers.

Baseline methods

To assess the performance of the proposed EMSTM, we compared it with several baseline methods. The baseline methods used were as follows:

- All_0: This method sets all prediction outputs to 0, effectively predicting no emotions.
- Binary relevance (BR) [56]: BR treats the multi-label classification problem as multiple binary classification problems, where each emotion label is predicted independently.
- Backpropagation for Multi-Label Learning (BP-MLL) [57]: BP-MLL captures multi-label features by optimizing an error function that incorporates pairwise label dependencies.
- Graph Memory Fusion Network (G-MFN) [52]: G-MFN combines a dynamic fusion graph and a memory fusion network with 3 parallel LSTMs to model multi-modal interactions.
- Transformer-Based Joint-Encoding (TBJE) [58]: TBJE integrates a mono-modal transformer and a multi-modal transformer using a joint-encoding method to capture the relationships between different modalities.
- Mu-Net [59]: Mu-Net utilizes context gated recurrent units (GRUs), state GRUs, and emotion GRUs to simulate different modalities and combines them through a pairwise attention mechanism.
- DEAN [60]: DEAN integrates a cross-modal transformer, multi-modal BiLSTM system, and gating block to emulate human emotional arousal processes, excelling in both multi-modal sentiment analysis and emotion recognition.
- FDMER [61]: FDMER utilizes common and private encoders to create modality-invariant and modality-specific subspaces, guided by a modality discriminator and specialized loss functions, to achieve refined multi-modal emotion recognition.
- MCR [62]: MCR employs specialized units for target and source modalities to facilitate nuanced cross-modal interaction and fusion, thereby enhancing the ability of the model to understand emotions.

The baseline methods serve as a comparative framework for evaluating the performance of the proposed EMSTM on the CMU-MOSEI dataset. It is crucial to note that the train/test splits used for these baseline methods were consistent with those used for the proposed EMSTM, ensuring a fair and direct comparison.

Training details

In the proposed EMSTM, we utilized a set of multi-head cross-attention blocks. Specifically, we used 6 heads and set the number of hidden dimensions to 96. We employed the AdamW optimizer [63] for training with a learning rate of 1e−3 and a batch size of 64. During training, if the loss value on the validation set did not decrease, we applied a decay rate of 0.1 to reduce the learning rate twice and then stopped the training early.

To optimize the model parameters from a unified perspective and encourage similarity within each category while minimizing the similarity between categories, we drew inspiration from circle loss [64]. We leveraged the properties of LogSumExp and maximized the similarity within each emotion category. To make predictions after the model was trained, we searched for a boundary threshold *th* near 0 using a validation set. If the output result for a particular emotion was greater than *h*, we considered the emotion to be expressed. For a detailed comparison of the number of parameters in our model with those in the baseline methods, please refer to the Discussion section.

Results

The experimental results of the different methods for the CMU-MOSEI test set are presented in Table 1. We compared our proposed model with the traditional multi-label classification methods, BP-MLL and BR. The results demonstrate that the proposed model outperforms these methods.

The "W/O MSTN" entry in Table 1 represents results obtained by calculating multi-modal EEE without using the MSTN module. It can be observed that the results obtained by this method are comparable to, or even better than, the results obtained using baseline methods such as Mu-Net, which incorporates previous content.

The "EMSTM" entry corresponds to the results of our proposed method, where multi-modal EEE is used for mental state transition. The improvement achieved by EMSTM demonstrates the effectiveness of the proposed method. Substantial improvements were observed in the emotions that constitute a large proportion of the dataset. Notably, we achieved a leading score of 72.5/72.4 for the happiness emotion, surpassing the previously published best score of 70.0/68.4. However, the results for surprise and fear emotions, which account for a smaller proportion of the dataset, are closer to the "All_0" baseline. The weighted accuracy increased from 77.2 to 78.0 after adding the MSTN module. This indicates that the MSTN module can effectively track mental state.

Additionally, we examined the impact of the underlying feature changes. In the "EMSTM(L)" entry, we replaced the language features used in the proposed method with 768-dimensional features extracted by the pre-trained RoBERTa (base) model [65]. The results further improved compared with EMSTM when replacing the open-source input features with features extracted by us.

Results on Ren-CECps

Dataset and evaluation metrics

Further experiments were conducted using the Ren-CECps dataset to validate the effectiveness of the proposed model. Among the available multi-label emotional classification datasets, Ren-CECps is one of the few datasets with a suitable data volume.

The Ren-CECps dataset, comprising Chinese blog texts annotated for 8 emotions, served as another test bed for our model. The dataset includes 27,091 sentences for training and 7,681

Table 1. Performance results on the CMU-MOSEI dataset, with values in boldface indicating the best results for clarity. Higher values indicate better performance for all metrics. Weighted Acc is for 6 emotions. The dagger (†) includes Neutral.

	Happiness		Sadness		Anger		Fear		Disgust		Surprise		Weighted Acc
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	
All_0	46.3	29.3	75.8	65.3	77.0	67.0	**91.7**	87.8	82.7	74.9	90.5	86.0	67.9
BR	70.7	70.4	75.8	65.3	77.1	67.3	**91.7**	87.8	82.7	74.9	90.5	86.0	77.4
BP-MLL	65.4	61.9	71.4	71.9	77.6	73.8	**91.7**	87.8	82.5	81.3	90.5	86.0	74.6
G-MFN	66.3	66.3	60.4	66.9	62.6	72.8	62.0	**89.9**	69.1	76.6	53.7	85.5	63.5
TBJE	66.0	65.5	73.9	67.9	81.9	76.0	89.2	87.2	86.5	84.5	**90.6**	86.1	76.0
Mu-Net	70.0	68.4	**76.1**	**74.5**	**83.0**	**80.9**	89.7	87.0	**90.3**	**87.3**	87.4	84.0	78.3
DEAN	–	–	–	–	–	–	–	–	–	–	–	–	52.3†
FDMER	–	–	–	–	–	–	–	–	–	–	–	–	54.1†
MCR	–	–	–	–	–	–	–	–	–	–	–	–	55.2†
W/O MSTN	70.8	70.6	73.7	71.2	77.3	73.5	91.5	87.9	84.0	82.1	**90.6**	86.1	77.2
EMSTM	72.5	72.4	74.0	73.3	77.4	75.7	**91.7**	87.8	84.0	83.2	**90.6**	**86.2**	78.0
EMSTM(L)	**74.4**	**74.3**	74.6	74.0	79.1	78.0	**91.7**	87.8	85.4	84.5	90.4	**86.2**	**79.2**

sentences for testing. As with the CMU-MOSEI dataset, the original study ([66]) did not clarify whether a cross-subject paradigm was employed for data splitting, and the anonymized nature of the dataset precluded us from verifying this. We followed the original guidelines for data splitting and maintained consistency with baseline methods. This monomodal dataset was used to evaluate the performance of the MSTN module. For further details, refer to Li et al. [66].

For comparison with published state-of-the-art models, we adopted the following evaluation metrics: micro F1 score, macro F1 score, average precision (AP), coverage error (CE), and ranking loss (RL). These metrics differ from those used in the CMU-MOSEI experiment. The calculation codes for these metrics are available in the sklearn library. CE measures the error in the sampling frame, AP represents the average precision on the precision–recall curve, and RL quantifies the relative distance between input samples. The higher the micro F1 and macro F1 scores and AP, the better the performance, whereas lower values of CE and RL indicate better performance. Table 2 presents the results of using different methods with the Ren-CECps dataset.

Baseline methods

In addition to the BR and BP-MLL methods used in the CMU-MOSEI experiment, we included several other baseline methods for comparison with the proposed model on the Ren-CECps dataset.

- Classifier chain (CC) [67] considers the label dependencies by using the BR method.
- Label powerset (LP) [68] treats each label combination as a separate category.
- Deep pyramid CNN (DPCNN) [69] is a word-level network that captures global information.
- Hierarchical attention network (HAN) [70] constructs hierarchical structures to capture information.
- Multi-label emotion detection architecture from sentences (MEDA-FS) [71] is a multi-channel hierarchical model that achieved the best published results.
- RoBERTa-Multi-Attention (RoBERTa-MA) [72] employs the RoBERTa transformer architecture, enhanced with multiple attention mechanisms, to focus on salient input features for each emotion label in multi-label emotion classification.

Training details

We utilized a pretrained BERT model (base-Chinese) [73,74] to extract sentence features. Considering that the MSTN module operates on sentence-level features, we applied the same dimensionality reduction method as the multi-modal EEE calculation model, which involves pooling the maximum and average values of the feature sequence and concatenating them with the [CLS] feature that contains sentence information. Notably, the published model only uses the [*hrmCLS*] feature as the sentence-level feature. The optimization and training strategies were consistent with those used in the CMU-MOSEI experiment. For a detailed comparison of the number of parameters in our model with those in the baseline methods, please refer to the Discussion section.

Results

Table 2 shows the experimental results on the Ren-CECps dataset, revealing the varying impacts of different representation methods on performance. Models with hierarchical structures generally outperformed those without them. Our implemented FC model, consisting solely of a fully connected layer for emotion prediction based on language features, achieved results comparable to those of the best published model, MEDA-FS, even without additional techniques. Our proposed MSTN method capitalizes on contextual information from preceding text, thereby enhancing information transfer at the sentence

Table 2. Performance results on the Ren-CECps dataset, with values in boldface indicating the best results for clarity. Higher values indicate better performance for the Micro F1, Macro F1, and AP metrics, while lower values indicate better performance for the CE and RL metrics.

	Micro F1	Macro F1	AP	CE	RL
BR	46.40	34.79	63.69	2.8313	0.1789
CC	46.97	33.62	63.16	2.9721	0.1965
LP	45.15	42.51	62.62	2.9117	0.1861
BP-MLL	48.89	38.13	55.45	3.1272	0.3234
DPCNN	49.99	35.47	65.43	3.0555	0.1993
HAN	54.54	41.36	70.65	2.4631	0.1362
MEDA-FS	60.76	48.31	76.51	2.2226	0.1062
RoBERTa-MA	49.70	42.40	–	–	–
FC	57.57	42.91	77.43	2.1168	0.1144
MSTN	**64.42**	**48.76**	**81.32**	**1.9346**	**0.0898**

level. Ablation experiments confirmed that the introduction of the MSTN substantially improved performance across all evaluation metrics. Importantly, the train/test splits for these baseline methods were consistent with those for MSTN, ensuring a balanced comparison. The results further validate the effectiveness of MSTN, as it surpassed state-of-the-art published outcomes.

Discussion

Numerous existing multi-modal emotion recognition models that consider the previous context utilize RNNs such as GRUs to handle sequential features. In contrast, the proposed EMSTM approach divides the process into 2 parts: calculating EEE using pattern recognition methods and influencing mental state transitions using the proposed MSTN module. Based on the theory we adopted, emotional transitions depend on historical emotions and the influence of EEE. We explored this process using the CMU-MOSEI and Ren-CECps datasets, and the experimental results demonstrate that our new emotion transition paradigm aligns more closely with the process of human mental activity.

Our method offers several advantages over other models. First, in the temporal sequence, we dynamically transitioned emotions rather than transferring multi-modal features, resulting in a substantial reduction in the number of model parameters. Second, were placed the RNNs with parallel attention blocks at each time step, reducing the computational time required for the model. These design choices enhanced the efficiency and effectiveness of the proposed approach.

To quantitatively assess the efficiency of the proposed EMSTM, we compared the number of parameters for EMSTM and various baseline methods, as shown in Fig. 7. For models such as DEAN and FDMER, where the number of parameters is neither explicitly reported in the original publications nor available in open-source repositories, we estimated the figures utilizing the model size estimation methodology proposed by Kamper et al. [75]. Remarkably, our model requires a mere 0.59M parameters, which is not only considerably fewer than most baseline models but also competitive with models such as G-MFN (0.592M) and MCR (0.79M). This lean parameterization does not compromise performance; rather, it complements

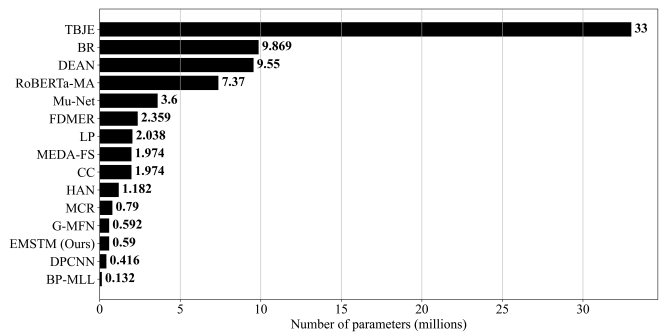

Fig. 7. Comparison of the number of model parameters for EMSTM and baseline methods.

the robustness and effectiveness of the proposed model, as evidenced by the superior results achieved on the CMU-MOSEI and Ren-CECps datasets.

About external emotional energy calculation

Observable behaviors can reflect EEE, which, in turn, affects mental state transitions. In our experiments, the pattern recognition model obtained EEE rather than directly inferring real emotions.

The comparison between "W/O MSTN" and "EMSTM" in Table 1 suggests that if emotions are directly inferred from EEE, similar to pattern recognition, comparable results can be achieved. This demonstrates that EEE has a substantial influence on the current mental state.

The module we adopted utilized only current features from different modalities and did not involve previous content. To calculate character-level features, we employed a parallel attention mechanism, which has been proven to be much faster than recurrent mechanisms in terms of training and inference speed.

About the MSTN

In terms of the process of mental state transition, EEE can only affect changes in emotions and does not directly determine emotions. This was more effective and closer to our intuition for expressing emotions using the MSTN module.

The CMU-MOSEI dataset exhibits a skip sentence phenomenon with cases of partial monologues and some unlabeled sentences, indicating that the data are not completely coherent in terms of context. This slight inconsistency with the assumptions of our model resulted in a small improvement when adding MSTN. The F1 scores for several emotions increased by approximately 2%, as shown in Table 1.

However, the Ren-CECps dataset, which lacked the skip sentence phenomenon, was closer to the real world. The results of "FC" vs. "MSTN" demonstrate that simply adding MSTN can increase F1 scores by more than 6%. This confirms that simulating human mental state transitions in emotion inference is closer to real emotions than pattern recognition alone.

About the input modality feature

People exhibit different external emotional behaviors in various situations; therefore, the modalities we adopt may vary. A multimodal model should be insensitive to modality redundancy or missing modalities. The proposed EMSTM does not have fixed-feature input-format requirements. To clarify how EMSTM handles different numbers of modal inputs, the model employs a multi-head cross-attention mechanism, as detailed in the "Cross-modality feature fusion" section. This mechanism allows the dynamic fusion of features across various modalities, thereby rendering the model highly scalable. In our experiments on both monomodal and multi-modal datasets, we demonstrated that this approach enables EMSTM to effectively adapt to different numbers of modal inputs. The modal-fusion mechanism in the calculation process can be easily extended to a different number of modalities. In our experiments with these 2 datasets, we used both multi-modal and monomodal inputs to verify the effectiveness of our model. Furthermore, the feature shapes of the different modalities may vary. These results demonstrate that EMSTM exhibits good scalability and generalization ability.

Better feature representation typically leads to improved results. For example, features extracted by large-scale pre-trained models are generally considered superior to GloVe features in capturing semantic expressions. The results of "EMSTM" vs. "EMSTM(L)" in Table 1 demonstrate that replacing GloVe features with RoBERTa features slightly improves the F1 scores. For the Ren-CECps dataset, we adopted the same feature extraction method as MEDA-FS. Moreover, we concatenated the sentence-level mean and maximum features with the [CLS] token, enriching the semantic information of the features. With only the classification layer added, the performance of our model was comparable to that of the state-of-the-art published models.

About data bias

We initially posited that EMSTM is a highly versatile model capable of processing both monomodal and multi-modal data. To substantiate this, we conducted a psychological questionnaire survey involving 200 participants. We used scientific sampling methods to obtain the general parameters, denoted by θ, for the emotional transition matrix T [30]. While these parameters serve as a robust starting point, it is crucial to specify that this versatility pertains to its capability to process different types of data modalities rather than its ability to adapt to varying data distributions. Therefore, fine-tuning the model for specific applications is recommended to achieve optimal performance. In our MSTN experiments, we used $0a$ as the initial value and adaptively adjusted the parameters to better fit the characteristics of the different datasets.

However, the datasets used were inevitably biased. For example, Fig. 6 shows that the number of positive emotions, such as happiness, is much greater than that of negative emotions. This caused a large proportion of the data to have a heavy weight during training, resulting in biased results. In the CMU-MOSEI dataset, the results for fear and surprise were similar to the predictions of the All_0 baseline in both our model and other models. We also attempted data augmentation methods such as linear interpolation, but the results did not differ substantially.

About multi-label classification

Given that multiple emotions may coexist simultaneously, we treated emotional recognition as a multi-label classification problem. Existing methods, such as BP-MLL and BR, are commonly used for multi-label classification; however, they do not explicitly consider the relationships among emotion categories. In emotion recognition, the probability of co-occurrence between positive emotions was higher than that between positive and negative emotions. Therefore, BP-MLL and BR may not perform well in capturing EEE.

To address this issue, we extended the circle-loss method to multi-label classification. Unlike BR, we did not use the sigmoid activation function for classification. Instead, the output results were distributed as real numbers and were not constrained to originate from 0. During prediction, we determined a threshold close to 0 through the validation set and used it to determine whether a certain emotion was expressed. In Tables 1 and 2, the proposed method substantially outperforms the commonly used BR and BP-MLL methods in terms of EEE calculation.

Conclusion

In this study, we proposed a new paradigm for multi-modal emotional transitions based on the theory of emotional recognition: EMSTM. The proposed approach involves calculating EEE based on multi-modal features and simulating the emotional transition process using MSTN, which is influenced by EEE. The experimental results on 2 multi-label emotion recognition datasets, CMU-MOSEI and Ren-CECps, demonstrated that EMSTM outperforms the best published results and confirmed its effectiveness. In addition to achieving high accuracy, the proposed model has the advantage of having fewer parameters, which makes it suitable for deployment on mobile devices and robots. This study also evaluated the general mental state transition probability and the impact of EEE from a broader perspective.

In our future work, we plan to consider emotional personality, as the impact of EEE varies from person to person. For example, calmer individuals may be less affected by EEE. We will explore personalized emotional transition models based on the emotional fluctuations of each individual to provide support for emotional monitoring and guidance.

Acknowledgments

Author contributions: F.-J.R. conceived the idea. X.K. supervised the study. Y.-Y.Z. and J.-W.D. completed the construction of the ESTM algorithm, and D.F., T.-H.S., Z.L., and Z-Y.J. completed the MSTN algorithm. K.M., T.-H.L., and S.N. completed the statistical analyses. Y.-Y.Z., J.-W.D., and X.K. wrote and revised the manuscript. X.K. contributed to the polishing of the manuscript.

Competing interests: The authors declare that they have no competing interests.

Data Availability

The code for our model is available at https://github.com/youngzhou97qz/Multimodal-emotion-processing.

References

1. Ren F, Bao Y. A review on human-computer interaction and intelligent robots. *Int J Inf Technol Decis Mak*. 2020;19(01):5–47.
2. Deng J, Ren F. A survey of textual emotion recognition and its challenges. *IEEE Trans Affect Comput*. 2021;14(1):49–67.
3. Majumder N, Hong P, Peng S, Lu J, Ghosal D, Gelbukh A, Mihalcea R, Poria S. MIME: MIMicking emotions for empathetic response generation. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.01454
4. Newen A, Welpinghus A, Juckel G. Emotion recognition as pattern recognition: The relevance of perception. *Mind Lang*. 2015;30(2):187–208.
5. Zhang Y, Chen J, Liu B, Yang Y, Li H, Zheng X, Chen X, Ren T, Xiong N. COVID-19 public opinion and emotion monitoring system based on time series thermal new word mining. arXiv. 2020. https://doi.org/10.48550/arXiv.2005.11458
6. Satrio D, Priyanto SH, Nugraha AK. Viral marketing for cultural product: The role of emotion and cultural awareness to influence purchasing intention. *Monten J Econ*. 2020;16(2):77–91.
7. Ren F, Kang X, Quan C. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE J Biomed Health Inform*. 2015;20(5):1384–1396.
8. Sharma A, Choudhury M, Althoff T, Sharma A. Engagement patterns of peer-to-peer interactions on mental health platforms. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence. vol. 14; 2020. p. 614–625.
9. Ai H, Litman DJ, Forbes-Riley K, Rotaru M, Tetreault J, Purandare A. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In: *Ninth International Conference on Spoken Language Processing*. Pittsburgh (PA): International Speech Communication Association; 2006.
10. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2018;41(2):423–443.
11. Quan C, Ren F. Weighted high-order hidden Markov models for compound emotions recognition in text. *Inf Sci*. 2016;329:581–596.
12. Kang X, Ren F, Wu Y. Exploring latent semantic information for textual emotion recognition in blog articles. *IEEE CAA J Autom Sin*. 2017;5(1):204–216.
13. Chen T, Ju S, Ren F, Fan M, Gu Y. EEG emotion recognition model based on the LIBSVM classifier. *Measurement*. 2020;164:Article 108047.
14. Dong Y, Ren F. Multi-reservoirs EEG signal feature sensing and recognition method based on generative adversarial networks. *Comput Commun*. 2020;164:177–184.
15. Xu G, Li W, Liu J. A social emotion classification approach using multi-model fusion. *Futur Gener Comput Syst*. 2020;102:347–356.
16. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R. ICON: Interactive conversational memory network for multimodal emotion detection. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels (Belgium): Association for Computational Linguistics; 2018. p. 2594–2604.
17. Kuppens P, Verduyn P. Emotion dynamics. *Curr Opin Psychol*. 2017;17:22–26.
18. Hall CS. *A primer of Freudian psychology*. Cleveland (OH): Pickle Partners Publishing; 2016.
19. Xiaolan P, Lun X, Xin L, Zhiliang W. Emotional state transition model based on stimulus and personality characteristics. *China Commun*. 2013;10(6):146–155.
20. Stocker E, Seiler R, Schmid J, Englert C. Hold your strength! Motivation, attention, and emotion as potential psychological mediators between cognitive and physical self-control. *Sport Exerc Perform Psychol*. 2020;9(2):167.
21. Ren F. Affective information processing and recognizing human emotion. *Electron Notes Theor Comput Sci*. 2009;225:39–50.
22. Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf Fusion*. 2020;59:103–126.
23. Jiang Y, Li W, Hossain MS, Chen M, Alelaiwi A, Al-Hammadi M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf Fusion*. 2020;53:209–221.
24. D'mello SK, Kory J. A review and meta-analysis of multimodal affect detection systems. *ACM Comput Surv*. 2015;47(3):1–36.
25. Bhaskar J, Sruthi K, Nedungadi P. Hybrid approach for emotion classification of audio conversation based on text and speech mining. *Procedia Comput Sci*. 2015;46:635–643.
26. Huang Y, Yang J, Liao P, Pan J. Fusion of facial expressions and EEG for multimodal emotion recognition. *Comput Intell Neurosci*. 2017;2017:2107451.
27. Sun B, Li L, Zhou G, Wu X, He J, Yu L, Li D, Wei Q. Combining multimodal features within a fusion network for emotion recognition in the wild. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. Seattle (WA): Association for Computing Machinery; 2015. p. 497–502.
28. Corchs S, Fersini E, Gasparini F. Ensemble learning on visual and textual data for social image emotion classification. *Int J Mach Learn Cybern*. 2019;10(8):2057–2070.
29. Yang L, Hong fL, Guo W. Text based emotion transformation analysis. *Comput Eng Sci*. 2011;9:026.
30. Xiang H, Jiang P, Xiao S, Ren F, Kuroiwa S. A model of mental state transition network. *IEEJ Trans Electron Inf Syst*. 2007;127(3):434–442.
31. Teoh T-T, Cho S-Y. Notice of retraction: Human emotional states modeling by hidden Markov model. Paper presented at: 2011 Seventh International Conference on Natural Computation; 2011 Jul 26–28; Shanghai, China.
32. Sun X, Pei Z, Zhang C, Li G, Tao J. Design and analysis of a human-machine interaction system for researching Human's dynamic emotion. *IEEE Trans Syst Man Cybern Syst*. 2019;51(10):6111–6121.
33. Winata GI, Madotto A, Lin Z, Shin J, Xu Y, Xu P, Fung P. CAiRE—HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification. arXiv. 2019. https://doi.org/10.48550/arXiv.1906.04041
34. Zahiri SM, Choi JD. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In: *Workshops at the Thirty-Second AAAI Conference*

on Artificial Intelligence. New Orleans (LA): Association for the Advancement of Artificial Intelligence; 2018.
35. Bae S, Choi J, Lee S-G. SNU_IDS at SemEval-2019 task 3: Addressing training-test class distribution mismatch in conversational classification. arXiv. 2019. https://doi.org/10.48550/arXiv.1903.02163
36. Al Chanti DA, Caplier A. Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. *IEEE Trans Affect Comput*. 2018;12(2):363–376.
37. Sun X, Xia P, Ren F. Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition. *Neurocomputing*. 2021;444:378–389.
38. Wu M, Su W, Chen L, Pedrycz W, Hirota K. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition. *IEEE Trans Affect Comput*. 2020;13(2):805–817.
39. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialogueRNN: An attentive RNN for emotion detection in conversations. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu (HI): Association for the Advancement of Artificial Intelligence. vol. 33; 2019. p. 6818–6825.
40. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv. 2014. https://doi.org/10.48550/arXiv.1412.3555
41. Ren F, Wang Y, Quan C. A novel factored POMDP model for affective dialogue management. *J Intell Fuzzy Syst*. 2016;31(1):127–136.
42. Zhou Y, Ren F. CERG: Chinese emotional response generator with retrieval method. *Research*. 2020;2020:2616410.
43. Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. arXiv. 2019. https://doi.org/10.48550/arXiv.1908.11540
44. Wang Y, Zhang J, Ma J, Wang S, Xiao J. Contextualized emotion recognition in conversation as sequence tagging. In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics; 2020. p. 186–195.
45. Canestri J. *Emotion in the psychoanalytic theory*Oxford: Oxford University Press; 2012.
46. Kang X, Shi X, Wu Y, Ren F. Active learning with complementary sampling for instructing class-biased multi-label text emotion classification. *IEEE Trans Affect Comput*. 2020;14(1):523–536.
47. Zhou Y, Kang X, Ren F. Prompt consistency for multi-label textual emotion detection. *IEEE Trans Affect Comput*. 2023;15(1):121–129.
48. Ren F, Liu Z, Kang X. An efficient framework for constructing speech emotion corpus based on integrated active learning strategies. *IEEE Trans Affect Comput*. 2022;13(4):1929–1940.
49. Kang X, Ren F, Wu Y. Semisupervised learning of author-specific emotions in micro-blogs. *IEEJ Trans Electr Electron Eng*. 2016;11(6):768–775.
50. Ren F, Kang X. Employing hierarchical Bayesian networks in simple and complex emotion topic analysis. *Comput Speech Lang*. 2013;27(4):943–968.
51. Kang X, Ren F. Predicting complex word emotions and topics through a hierarchical Bayesian network. *China Commun*. 2012;9(3):99–109.
52. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. 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)*. Melbourne (Australia): Association for Computational Linguistics; 2018. p. 2236–2246.
53. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L. Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv. 2017. https://doi.org/10.48550/arXiv.1706.03762
54. Wu Y, Chen Y, Wang L, Ye Y, Liu Z, Guo Y, Fu Y. Large scale incremental learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2019. p. 374–382.
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Muller A, Nothman J, Louppea G, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
56. Luaces O, Díez J, Barranquero J, del Coz JJ, Bahamonde A. Binary relevance efficacy for multilabel classification. *Prog Artif Intell*. 2012;1(4):303–313.
57. Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans Knowl Data Eng*. 2006;18(10):1338–1351.
58. Delbrouck J-B, Tits N, Brousmiche M, Dupont S. A transformer-based joint-encoding for emotion recognition and sentiment analysis. arXiv. 2020. https://doi.org/10.48550/arXiv.2006.15955
59. Shenoy A, Sardana A. Multilogue-net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation. arXiv. 2020. https://doi.org/10.48550/arXiv.2002.08267
60. Zhang F, Li XC, Lim CP, Hua Q, Dong CR, Zhai JH. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. *Inf Fusion*. 2022;88:296–304.
61. Yang D, Huang S, Kuang H, Du Y, Zhang L. Disentangled representation learning for multimodal emotion recognition. In: *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa (Portugal): Association for Computing Machinery; 2022. p. 1642–1651.
62. Yang D, Liu Y, Huang C, Li M, Zhao X, Wang Y, Yang K, Wang Y, Zhai P, Zhang L. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowl-Based Syst*. 2023;265:Article 110370.
63. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv. 2017. https://doi.org/10.48550/arXiv.1711.05101
64. Sun Y, Cheng C, Zhang Y, Zhang C; Zheng L; Wang Z; Wei Y. Circle loss: A unified perspective of pair similarity optimization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2020. p. 6398–6407.
65. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized bert pretraining approach. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.11692
66. Li J, Ren F. Creating a Chinese emotion lexicon based on corpus Ren-CECps. In: *Proceedings of the 2011 IEEE International Conference on Cloud Computing and Intelligence Systems*. Beijing (China): IEEE; 2011. p. 80–84.
67. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. *Mach Learn*. 2011;85(3):333.
68. Read J, Puurula A, Bifet A. Multi-label classification with meta-labels. In: *Proceedings of the 2014 IEEE International Conference on Data Mining*. Shenzhen (China): IEEE; 2014. p. 941–946.

69. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver (Canada): Association for Computational Linguistics; 2017. p. 562–570.
70. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego (CA): Association for Computational Linguistics; 2016. p. 1480–1489.
71. Deng J, Ren F. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Trans Affect Comput*. 2020;14(1): 475–486.
72. Ameer I, Bölücü N, Siddiqui MHF, Can B, Sidorov G, Gelbukh A. Multi-label emotion classification in texts using transfer learning. *Expert Syst Appl*. 2023;213:Article 118534.
73. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. https://doi.org/10.48550/arXiv.1810.04805
74. Wolf T, Chaumond J, Debut L, Sanh V, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020. p. 38–45.
75. Kamper T, Wegerhoff M, Brücher H. Making modal analysis easy and more reliable—Reference points identification and model size estimation. In: *Proceedings of the 13th Aachen Acoustics Colloquium*. Aachen (Germany): HEAD acoustics GmbH; 2022. p. 1–10.