ND-MRM: Neuronal Diversity Inspired Multisensory Recognition Model

Qixin Wang, Chaoqiong Fan, Tianyuan Jia, Yuyang Han, Xia Wu*

School of Artificial Intelligence, Beijing Normal University, Beijing, China
{qxwang, fcq, tianyj, yuyang_han}@mail.bnu.edu.cn, wuxia@bnu.edu.cn

Abstract

Cross-sensory interaction is a key aspect of multisensory recognition. Without cross-sensory interaction, artificial neural networks show inferior performance in multisensory recognition. On the contrary, the human brain has an inherently remarkable ability in multisensory recognition, which stems from the diverse neurons that exhibit distinct responses to sensory inputs, especially the multisensory neurons with multisensory responses hence enabling cross-sensory interaction. Based on this neuronal diversity, we propose a Neuronal Diversity inspired Multisensory Recognition Model (ND-MRM), which, similar to the brain, comprises unisensory neurons and multisensory neurons. To reflect the different response characteristics of diverse neurons in the brain, special connection constraints are innovatively designed to regulate the feature transmission in the ND-MRM. Leveraging this novel concept of neuronal diversity, our model is biologically plausible, enabling more effective recognition of multisensory information. To validate the performance of the proposed ND-MRM, we employ a multisensory emotion recognition task as a case study. The results demonstrate that our model surpasses state-of-the-art brain-inspired baselines on two datasets, proving the potential of brain-inspired methods for advancing multisensory interaction and recognition.

Introduction

Compared to unisensory recognition, multisensory recognition, which considers information from various senses, demonstrates superior performance. This superiority has led to extensive research in multisensory recognition tasks, resulting in significant progress in areas such as emotion recognition (Ju et al. 2020; Zhang et al. 2022), medical diagnosis (Boehm et al. 2022a,b), and intelligent robotics (Papanastasiou et al. 2019; Heredia et al. 2022).

Research in multisensory recognition mainly focuses on the fusion of features from multiple senses. Various approaches, such as CNN-based, RNN-based, and their hybrid with other algorithms, have been utilized to realize the integration, demonstrating promising results in multisensory recognition tasks (Zhang et al. 2017b; Gong et al. 2021; Zhang et al. 2021; Gültekin et al. 2022). Additionally, deep generative models, including variational autoencoders (VAEs) (Lee and Pavlovic 2021; Wang et al. 2022), generative adversarial networks (GAN) (Zhang et al. 2017a; Kosaraju et al. 2019; Zhu et al. 2023), and so on, also show their excellent performance because of their superiority in feature representation (Suzuki and Matsuo 2022). However, many of these models do not consider the interaction between different senses, a crucial aspect for effective multisensory recognition (Mansouri-Benssassi and Ye 2020). In contrast, transformer-based cross-modal models, such as MulT (Bhattacharjee et al. 2022), achieve the interaction between different senses based on diverse attention mechanisms. Nevertheless, some concerns regarding their reliance on a large amount of computation and large datasets remain, leading to low computational efficiency and high training costs (Shin, Ishii, and Narihira 2022). With the limitations of the previous research, there is an urgency for a multisensory recognition model that can facilitate intricate cross-sensory interaction without excessive computation and data requirements, enabling more efficient recognition of multisensory information.

The human brain possesses an inherently remarkable ability to effectively perceive and recognize the external environment by fully utilizing multisensory information, such as vision and hearing(McDonald, Teder-Salejarvi, and Ward 2001; Ohshiro, Angelaki, and DeAngelis 2011; Stein, Stanford, and Rowland 2014). Neuroscience research has demonstrated that this ability of the brain stems from diverse neurons, including unisensory neurons for each sense and multisensory neurons (Alvarado et al. 2007). These neurons exhibit distinct responses to sensory inputs, where unisensory neurons respond only to single sensory information, while multisensory neurons respond to multisensory information (Stein and Stanford 2008; Stevenson et al. 2014). This diversity in neuronal response characteristics enables the human brain to effectively capture cross-sensory interaction, leading to superior abilities in multisensory recognition (Laurienti et al. 2005; Holmes 2007).

Therefore, based on the facilitation of diverse neurons including unisensory neurons and multisensory neurons to multisensory integration, we propose a Neuronal Diversity inspired Multisensory Recognition Model (ND-MRM) through spiking neural networks (SNN). This model is more biologically plausible, incorporating the novel concept of neuronal diversity, thus enabling more effective recognition

*Corresponding author

of multisensory information. This research makes several important contributions, summarized as follows.

- Aligning with the neuronal diversity observed in the human brain, both unisensory neurons and multisensory neurons are comprised in the ND-MRM, aiming to model and learn the cross-sensory interaction of multisensory information.
- Special connection constraints are innovatively designed to regulate the feature transmission in the ND-MRM. This configuration sufficiently reflects the different response characteristics of diverse neurons.
- We conducted several experiments to evaluate the effectiveness of our model in a typical multisensory emotion recognition task. The results demonstrate consistently superior performance to the state-of-the-art brain-inspired baselines.

Related Work

Brain-inspired models have been proposed to effectively recognize information from multiple senses, particularly in the context of multisensory emotion recognition. This section provides a brief overview of such models.

The $Convergence$ model (Benssassi and Ye 2023) draws inspiration from the convergence theory of multisensory emotion recognition. According to this theory, information from individual senses converges in higher-order multisensory regions for fusion and recognition(Stein and Meredith 1993). The model incorporates a convergence layer designed to recognize concatenated features from visual and auditory senses, symbolizing higher-order multisensory regions. Preceding the convergence layer, two separate layers extract features from each sense. However, the absence of cross-sensory interaction limits the performance improvement of this model.

The $Enhancement$ model (Benssassi and Ye 2023) is inspired by the enhancement theory for multisensory recognition, where visual information significantly influences auditory cortex activity in the human brain (Molholm et al. 2002; Jessen and Kotz 2013). In this model, the auditory features extraction layer receives inputs not only from the auditory input layer but also from the visual layer. Unfortunately, the model only considers unidirectional connections from the visual sense to the auditory sense, failing to achieve cross-sensory interaction for emotion recognition.

Compared with these two models that ignore the cross-sensory interaction, the $Synch\text{-}graph$ model incorporates neural synchrony as a means to capture interaction in multisensory information (Mansouri-Benssassi and Ye 2020). Neural synchrony reflects simultaneous neural oscillations of different neuron groups connected by synapses, considered vital for multisensory interaction (Stein 2012). The model establishes bidirectional connections between auditory and visual neurons to realize interaction. However, a notable shortcoming is its diminished performance in the presence of disturbances in sensory information, such as noise.

In summary, while the first two models lack cross-sensory interactions for multisensory emotion recognition, the $Synch\text{-}graph$ model attempts to capture such interactions inspired by neural synchrony but exhibits poor robustness. This highlights the need for novel brain mechanisms to inspire the development of models capable of intricate cross-sensory interactions.

Neuronal Diversity

In the human brain, the efficient processing and recognition of multisensory information rely on diverse neurons with distinct responses to sensory inputs. This section introduces these neurons, delineates their unique response characteristics, and elucidates their role in facilitating cross-sensory interaction.

Key regions responsible for multisensory recognition include the Superior Colliculus (SC)(Cuppini et al. 2011), and the posterior superior temporal sulcus and gyrus (STS/STG) (Murray and Wallace 2011; Chabrol et al. 2015). Within these regions, both multisensory neurons and unisensory neurons are prevalent. Multisensory neurons respond to stimuli from more than one sense, while unisensory neurons respond exclusively to a single sense. Further classification of unisensory neurons distinguishes between those responsive to visual stimuli and those responsive to auditory stimuli (Stein and Stanford 2008; Stevenson et al. 2014).

Multisensory neurons exhibit a significantly heightened response to multisensory stimuli, surpassing the response to any single sense, particularly when stimuli share a common source (Cuppini, Magosso, and Ursino 2011). Conversely, unisensory neurons display no substantial changes between their response to multisensory stimuli and their response to single sensory stimuli. Due to these distinctive response characteristics, both unisensory and multisensory neurons can selectively activate the next neuron, influencing the type and intensity of activation. Consequently, information from different senses can interact and complement each other among these neurons, achieving cross-sensory interaction and facilitating multisensory recognition (Allman, Keniston, and Meredith 2009).

Proposed ND-MRM Model

Inspired by the aforementioned superiority of neuronal diversity in cross-sensory interaction, we propose our ND-MRM model, as shown in Figure 1. Specifically, it consists of unisensory neurons dedicated to visual and auditory senses, along with multisensory neurons. This parallels the diverse neurons observed in the human brain. Besides, special connection constraints are designed to regulate the feature transmission, reflecting the different response characteristics of diverse neurons in the brain. Therefore, our model is biologically plausible, enabling more effective recognition of multisensory information.

In the following, we first introduce the overall framework of the ND-MRM model in the first subsection. The cores of our model, which are the diverse neurons and connection constraints, are introduced in the next two subsections, respectively.

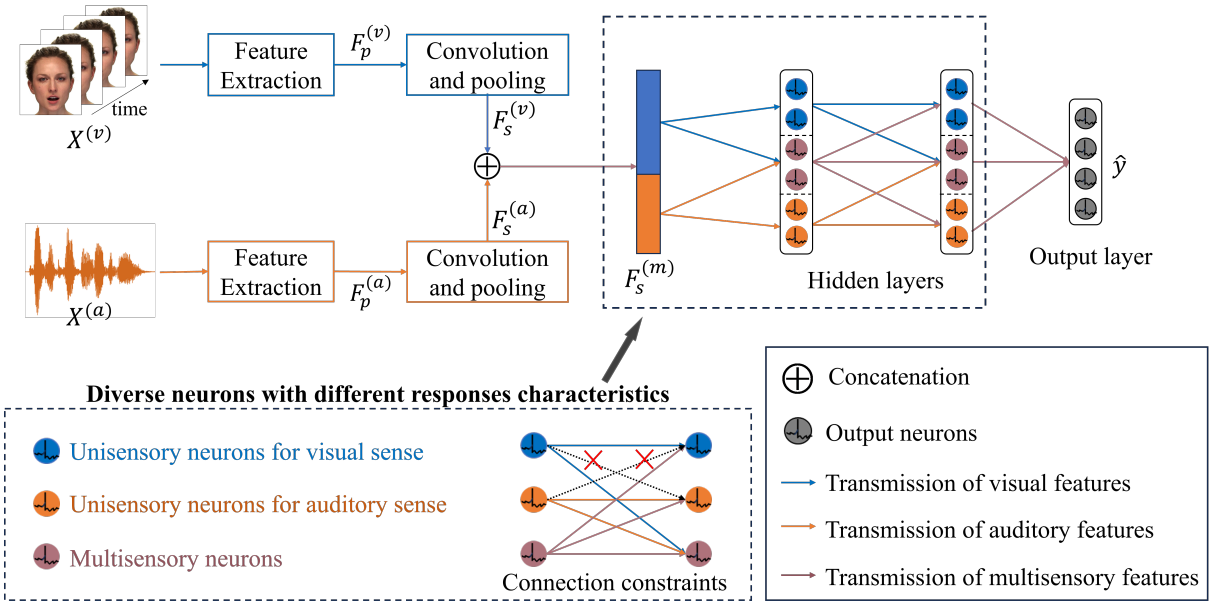

Figure 1: The overall architecture of the proposed ND-MRM model.

Overall Architecture

We consider two sensory modalities vision(v) and audio(a). The input data of these two senses are denoted as $X^{(v)} \in R^{T^{(v)} \times D^{(v)}}$ for each sample, respectively, where $T^{(\cdot)}$ and $D^{(\cdot)}$ are used to represent the time dimension and feature dimension of the input data. After preprocessing and feature extraction, the primary features of each modality, $F_p^{(v)}$, $F_p^{(a)}$, are obtained. Then they are further processed as semantic features $F_s^{(v)}$ and $F_s^{(a)}$ through two encoders. Each encoder is composed of a convolutional layer and a pooling layer by SNN.

After these processes, the multisensory features $F_s^{(m)}$ are concatenated from visual and auditory features $F_s^{(v)}, F_s^{(a)}$. In the ND-MRM, two hidden layers and an output layer are used to recognize the multisensory features. Specifically, three kinds of neurons are comprised of the two hidden layers, including unisensory neurons for visual sense, unisensory neurons for auditory sense, and multisensory neurons. Besides, special constraints are designed to restrict the connection between diverse neurons, regulating the feature transmission in these layers.

Based on the overall architecture of the ND-MRM, the classical leaky integrated-and-fire (LIF) model is taken as the neuron model. In addition, neuronal plasticity is also considered, which can improve the learning ability of the network (Jia et al. 2021). Therefore, the neuron model adds an adaptive firing threshold, which is set by the ordinary differential equation. For example, the update of the membrane potential of the i-th neuron in the l-th hidden layer is shown as,

$$C\frac{dV_i(t)}{dt} = g\left(V_i(t) - V_1\right)\left(1 - S_i(t)\right) - \gamma a_i(t) + \sum_{j=1}^{n_{l-1}} \left(W_l \odot mask_l\right)_{i,j} x_{j,l} \tag{1}$$

$$\begin{cases} V_i(t) = V_2, \\ S_i(t) = 1, \end{cases} if\left(V_i(t) = V_{th}\right), \tag{2}$$

$$\frac{da_i(t)}{dt} = (\alpha - 1)\, a_i(t) + \beta S_i(t) \tag{3}$$

where C is the capacitance parameter, g is the conductance value, $V_i(t)$ is the membrane potential of the i-th neuron at timing t, $S_i(t)$ is the firing flag, V_1 is the resting potential, V_2 is the reset membrane potentials, V_{th} is the firing threshold. W_l is the fully connected weight from the (l-1)-th layer to the l-th layer. The $mask_l$ represents the connection constraint from the (l-1)-th layer to the l-th layer, which is described in detail in the third subsection. the $\odot$ represents the Hadamard product. $(W_l \odot mask_l)_{i,j}$ is the weight from the j-th neuron in the (l-1)-th layer to the i-th neuron in the l-th layer. x_l is the input feature vector of the l-th hidden layer, and $x_{j,l}$ is the j-th dimension of x_l. The dynamic threshold $a_i(t)$ is accumulated during the period from the resetting to the membrane potential firing, and as the frequency of firing increases, the threshold also increases, and vice versa. α, β and γ are the hyperparameters.

Diverse Neurons

Since multisensory neurons and unisensory neurons for each sense are characterized by different responses to sensory stimuli, the type of neurons can be judged based on their spikes in SNN. This subsection first describes how to obtain the spikes of neurons, and then explains how to identify the

unisensory neurons for each sense and multisensory neurons based on their spikes.

Obtain spikes. To obtain spikes of neurons, we first establish two separate unisensory recognition models for visual and auditory sense with the same architecture. In the case of the unisensory recognition model for vision, the initial phase of this model involves visual feature extraction and feature encoding, mirroring the processes employed in the ND-MRM. The second phase of the unisensory recognition model is a network with two fully connected layers and an output layer. The recognition tasks in both unisensory recognition models are handled by their respective second phases, denoted as $Re^{(v)}, Re^{(a)}$ respectively.

After training these two unisensory recognition models, the spikes of neurons in the two fully connected layers are recorded, which are used to judge the type of neurons.

Identify diverse neurons. To highlight our ideas of how to identify the diverse neurons, we further explain the correspondence between the brain and the models, which include the ND-MRM as well as the two unisensory recognition models.

Processing unisensory information involves preprocessing, feature extraction, and recognition, culminating in deriving the final classification result. This comprehensive unisensory process aligns with the unisensory information recognition circuit observed in the human brain. Specifically, the extraction and encoding of unisensory features correspond to the primary unisensory pathway, akin to the ventral visual pathway. The entities $Re^{(v)}$ and $Re^{(a)}$ correspond to neuronal populations responsive to distinct senses within higher-order brain regions.

In the context of multisensory recognition, whether humans receive unisensory or multisensory information, it is ultimately recognized and analyzed by higher-order brain regions. The architectures of $Re^{(v)}$ and $Re^{(a)}$ closely align with the network structure comprising two hidden layers and an output layer in the ND-MRM, reflecting their shared capacity to emulate recognition functions in higher-order regions. However, distinctions persist in the neuronal populations they emulate. Due to the unique response characteristics of neurons, both unisensory neurons for the visual sense and multisensory neurons respond to visual stimuli, representing the neurons that $Re^{(v)}$ simulates. Similarly, $Re^{(a)}$ pertains to unisensory neurons for the auditory sense and multisensory neurons. Consequently, leveraging these divergent spike patterns of distinct neurons allows us to identify unisensory neurons for each sense and multisensory neurons, elaborated upon in the following.

The categorization of diverse neuron types is conducted in a layer-by-layer fashion. Illustrated in Figure 2, the spikes of neurons in the initial layer of $Re^{(v)}$ and $Re^{(a)}$ serve to identify neurons within the first hidden layer of the ND-MRM. This process extends similarly to the subsequent layer. The spike vectors, denoted as $s_{i,l}^{(v)}, s_{i,l}^{(a)}$, represent the spikes of all neurons in the l-th layer for the i-th sample of visual and auditory senses, respectively. Spike patterns, represented as $p_l^{(v)}$ and $p_l^{(a)}$, are derived through the averaging of spikes across the temporal scale and samples, formulated as follows:

$$\begin{aligned} p_l^{(v)} &= \frac{1}{N}\sum_{i=1}^{N} Mean_t\left(s_{i,l}^{(v)}\right) \\ p_l^{(a)} &= \frac{1}{N}\sum_{i=1}^{N} Mean_t\left(s_{i,l}^{(a)}\right) \end{aligned} \tag{4}$$

where $Mean_t(\cdot)$ represents the average value on timing scale. N is the total number of samples, $\frac{1}{N}$ represents to average spikes on samples. Then, the set of multisensory neurons is determined by,

$$M_l = Top\left(p_l^{(v)}, \rho\right) \cap Top\left(p_l^{(a)}, \rho\right) \tag{5}$$

where $Top(\cdot, \rho)$ is utilized to identify neurons exhibiting the highest firing levels based on their spike patterns, ρ serves as a hyperparameter within the range of 0 to 1. This parameter signifies our selection of neurons with the top ρ highest spikes, designating them as having the strongest firing levels. The multisensory neurons in the l-th layer, denoted as M_l, are determined as the intersection of visual neurons with the strongest firing levels and neurons associated with the auditory sense. Subsequently, the unisensory neurons in the l-th layer are discerned as the difference between all neurons and the multisensory neurons. This distinction is established through the expression,

$$\begin{aligned} U_l^{(v)} &= A_l^{(v)} - M_l \\ U_l^{(a)} &= A_l^{(a)} - M_l \end{aligned} \tag{6}$$

where the $U_l^{(v)}$ and $U_l^{(a)}$ represent unisensory neurons in the l-th layer, $A_l^{(v)}$ and $A_l^{(v)}$ represent all neurons in the l-th layer.

Connection Constraints

Throughout the process of multisensory recognition, unisensory neurons and multisensory neurons manifest distinct response characteristics. To capture these traits, specific connection constraints are devised to govern feature transmission within the ND-MRM, to facilitate cross-sensory interaction. This subsection provides a detailed description of the design and application of these constraints to the ND-MRM.

In the ND-MRM, the constraints are systematically designed on a layer-by-layer basis. Represented as weight masks, they are denoted as $mask_l \in R^{n_{l-1} \times n_l}$, where n_l signifies the number of neurons in the l-th hidden layer. These masks are matrices populated with 1 or 0, wherein 1 signifies connection and 0 indicates disconnection. The formulation of these masks is articulated as follows:

$$\begin{bmatrix} D\left(c_{1,l-1}, c_{1,l}\right) & \cdots & D\left(c_{1,l-1}, c_{n_l,l}\right) \\ \vdots & \ddots & \vdots \\ D\left(c_{n_{l-1},l-1}, c_{1,l}\right) & \cdots & D\left(c_{n_{l-1},l-1}, c_{n_l,l}\right) \end{bmatrix} \tag{7}$$

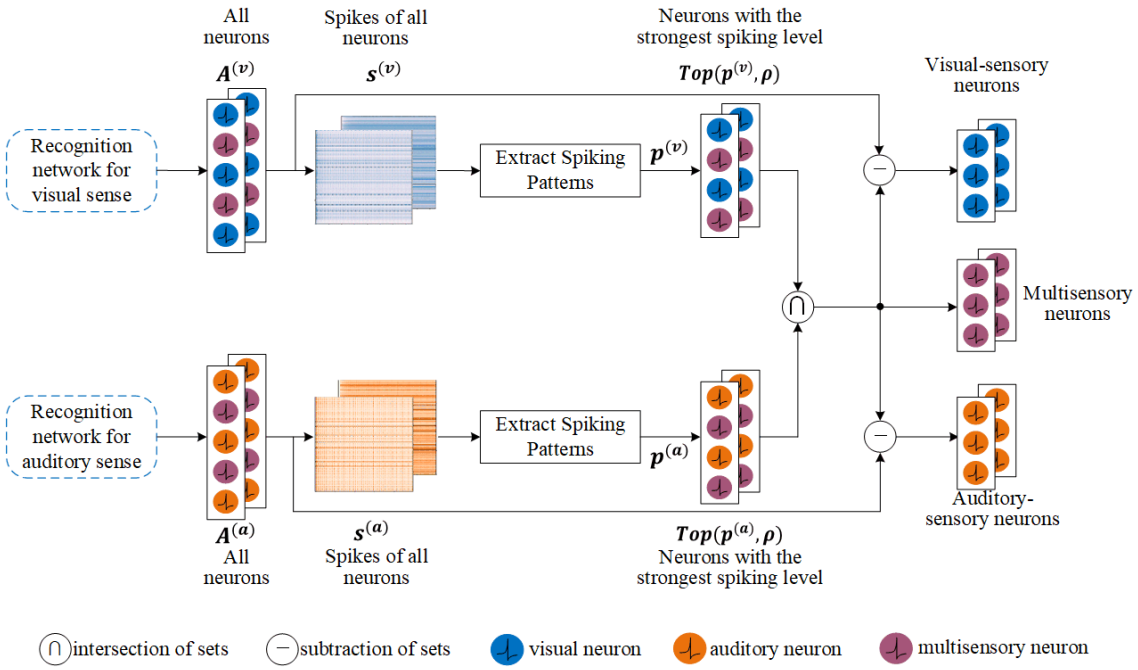

Figure 2: The categorization of diverse neuron types using their spikes.

where $D(\cdot)$ is a discriminant function and $c_{i,l}$ indicates the type of the i-th neuron in the l-th layer. The value of $D\left(c_{i,l-1}, c_{j,l}\right)$ is designed according to the following rules.

- If any two neurons in two adjacent layers are unisensory neurons of the same sense, the function value is set to 1.
- If at least one of any two neurons in two adjacent layers is a multisensory neuron, the function value is set to 1.
- Otherwise, it is set to 0.

By these constraints, the transmission of features between senses is facilitated, fostering cross-sensory interaction. To elucidate, the auditory features $F_s^{(a)}$ establish connections with the multisensory neurons of the first hidden layer M_1 and subsequently with the visual neuron of the second hidden layer $U_2^{(v)}$. This configuration allows for unidirectional transmission from the auditory sense to the visual sense. Analogously, a unidirectional transmission occurs from vision to audio. Consequently, cross-sensory interaction is effectively realized.

Case Study: Multisensory Emotion Recognition

In this section, we concentrate on the task of multisensory emotion recognition, using it as a case study. We perform a series of experiments to evaluate the performance of the ND-MRM on RAVDESS and eNTERFACE'05 datasets.

Experiments Setup

Datasets. Our model is evaluated using two datasets. The first dataset is eNTERFACE'05 (Martin et al. 2006), which comprises 42 participants, consisting of 81% male and 19% female participants. The audio is recorded at 48000Hz in 16-bit format, and the videos have a frame rate of 25 frames per second. Each participant expresses six distinct emotions: anger, disgust, fear, happiness, sadness, and surprise.

The second dataset utilized is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018). This dataset features a balanced gender distribution among its 24 participants, who are actors tasked with reading a sentence in eight distinct emotional states: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. In this study, our focus is specifically on the speech and video modalities within this dataset.

Feature extraction. Initially, the raw data needs to be extracted to serve as input features for the ND-MRM. For the visual modality, 15 frames are extracted at equal intervals for each video, denoted as $T(v)$, with a value of 15. Subsequently, facial contours are extracted and downscaled to a size of 28*28 to serve as visual features for each frame. Consequently, the final dimension of visual features $F_p^{(v)}$ for each video is $R^{15\times784}$.

In the auditory modality, the widely employed Mel-scale Frequency Cepstral Coefficients (MFCC) are extracted as auditory features $F_p^{(a)}$ for each speech. The average number of frames across all speeches is 280, and the feature dimension for each frame is 12. Consequently, the final dimension is $R^{12\times280}$.

Configurations. In the convolutional layers of the ND-MRM, the number of channels is set to 4, and the kernel size is $5 * 5$. Both $Re^{(v)}$ and $Re^{(v)}$ consist of two fully connected layers ($n_l = 200$) and an output layer with the same number of labels as each dataset.

Model	Neutral	Clam	Happy	Sad	Angry	Fearful	Disgust	Surprised	Weighted_Acc
$Convergence^*$	-	-	-	-	-	-	-	-	0.8130
$Enhancement^*$	-	-	-	-	-	-	-	-	0.7330
$Synch\text{-}Graph^*$			1.0000	**1.0000**	1.0000	0.9550	0.9310	0.9290	0.9830
$MR\text{-}SNN$	0.9655	0.9825	1.0000	0.9355	0.9310	0.9091	0.9474	0.9655	0.9537
$MulT^*$		-	-	-	-	-	-	-	0.7416
ND-MRM	**1.0000**	**1.0000**	**1.0000**	0.9933	**1.0000**	**0.9871**	**1.0000**	**0.9931**	**0.9963**
MRM	0.9292	0.9292	0.9690	0.9615	1.0000	0.8891	0.9315	0.9249	0.9418

Table 1: Comparison of accuracy for multisensory emotion recognition on RAVDESS dataset. * represents that the performance of this model is obtained from the relevant paper.

Model	Angry	Disgust	Fear	Happy	Sad	Surprised	Weighted_Acc
$Convergence^*$	-	-	-	-	-	-	0.8330
$Enhancement^*$	-	-	-	-	-	-	0.8330
$Synch\text{-}Graph^*$	0.9470	0.9550	**1.0000**	**1.0000**	0.9200	**1.0000**	0.9682
$MR\text{-}SNN$	0.9231	0.9180	0.9032	0.9118	0.9231	0.8955	0.9124
ND-MRM	**0.9814**	**1.0000**	0.9905	0.9905	**0.9235**	0.9818	**0.9767**
MRM	0.9232	0.9232	0.9456	0.9355	0.9032	0.9688	0.9333

Table 2: Comparison of accuracy for multisensory emotion recognition on eNTERFACE'05 dataset. * represents that the performance of this model is obtained from the relevant paper.

The hyperparameter ρ is initially set to 0.7. The capacitance C is $1\mu F/cm^2$, g is $0.2nS$, time constant is $1ms$, resting potential V_1 is equal to reset potential V_2 with $0mV$. The firing threshold is $0.5mV$ in the beginning. For the adaptive threshold, we set α=0.9, β=0.1, and γ=1.

Baselines. We select a set of brain-inspired methods, which have been described in detail in the previous section. In addition to these models, other models such as $MR\text{-}SNN$ and $MulT$ are also compared in this study.

- The **Convergence** (Benssassi and Ye 2023) model is inspired by the convergence theory with a convergence layer to recognize concatenated features of different senses.
- The **Enhancement** (Benssassi and Ye 2023) model is inspired by the enhancement theory with unidirectional interaction from visual neurons to auditory neurons.
- The **Synch-Graph** (Mansouri-Benssassi and Ye 2020) model is inspired by neural synchrony which is captured by a graph network and the recognition is achieved by GCN.
- The **MR-SNN** (Jia et al. 2022) model is proposed to recognize digits in MNIST and TIDigits datasets. It's based on the motifs which are topologies or connections between neurons and extracted from pre-trained networks that are used in unisensory tasks.
- The **MulT** (Tsai et al. 2019) adopts directional pairwise cross-modal attention, which attends to the interaction between multimodal sequences, to recognize multisensory information.

Overall Performance

The performance comparison between our proposed MD-MRM model and the state-of-the-art methods on the two datasets is shown in Table 1 and Table 2.

Our model exhibits superior performance on both datasets, achieving 99.63% on the RAVDESS dataset and 97.67% on the eNTERFACE'05 dataset. These results significantly surpass the state-of-the-art multisensory emotion recognition method for the same datasets.

Among the baselines, the best-performing state-of-the-art model on both datasets is the $Synch\text{-}Graph$ model, which has achieved 98.30% and 96.82%, respectively. This model learns synchrony patterns between audio and visual neuron groups. In contrast, our model recognizes concatenated features with inspiration from neuronal diversity, achieving cross-sensory interaction through connections between unisensory neurons and multisensory neurons, leading to efficient multisensory recognition.

Benefiting from these advantages, our model attains 99.63% on the RAVDESS dataset with eight classes, while the $Synch\text{-}Graph$ achieved 98.30% with six classes. The two classes not considered by the latter model are the labels of neutral and calm, where our model achieves 100% accuracy. Besides, our model outperforms the $Synch\text{-}Graph$ by 0.85% on the eNTERFACE'05 dataset.

Compared with other brain-inspired methods, such as the $Convergence$, $Enhancement$, and $MR\text{-}SNN$, our model outperforms them by 18.33%, 26.33%, 4.26% and 14.37%, 14.37%, 6.43% on the two datasets, respectively.

Furthermore, in addition to these brain-inspired methods, we compare the performance of $MulT$(Chumachenko, Iosifidis, and Gabbouj 2022) with our model specifically in the RAVDESS dataset. The accuracy achieved by MulT on

seven classes is 74.16%, and our model outperforms MulT by 25.47%.

In summary, these results underscore that our model excels on both datasets in the context of multisensory emotion recognition. The ND-MRM effectively captures interactions between different senses through neuronal diversity inspiration, demonstrating superiority over the pairwise cross-modal attention approach employed by $MulT$.

Ablation Studies

In this subsection, we conduct two ablation experiments to further investigate the effect of diverse neurons on multisensory recognition. Initially, we assess the performance of models featuring only one type of neuron responsive to all inputs. Subsequently, we explore the influence of the number of diverse neurons on the performance of the ND-MRM.

Ablation study on diverse neurons. In the ND-MRM, both multisensory neurons and unisensory neurons are incorporated to achieve cross-sensory interaction. A model containing only unisensory neurons cannot achieve cross-sensory interaction as unisensory neurons respond solely to single sensory stimuli. Therefore, we did not conduct an ablation experiment for the model comprising only unisensory neurons. The MRM, a model containing only multisensory neurons in the two hidden layers, is investigated in this ablation experiment.

The MRM achieves an accuracy of 94.18% on the RAVDESS dataset and 93.33% on the eNTERFACE'05 dataset. Table 1 and Table 2 show the accuracy of each emotion class between ND-MRM and MRM on these two datasets.

It can be seen that the ND-MRM outperforms the MRM by 5.45%, and 4.34% on the two datasets, respectively. Regarding the accuracy of each class, the performance of the MRM on each emotion class is significantly lower than that of the ND-MRM, except for the 'angry' class in the RAVDESS dataset. This is because the neurons in the two hidden layers are all multisensory. These experiments underscore the contribution of diverse neurons, encompassing both unisensory and multisensory neurons, to multisensory emotion recognition tasks.

Ablation study on the number of neurons. As described in the subsection of Diverse Neurons, the numbers of multisensory neurons and unisensory neurons are identified by the hyperparameter ρ. Therefore, we investigate how the number of diverse neurons influences the performance of the ND-MRM by varying ρ.

Table 3 shows the value of ρ, the corresponding number of multisensory neurons (nom), the proportion of multisensory neurons among all neurons (p), and the weighted accuracy (Acc) on two datasets.

Observing the table, as the hyperparameter ρ approaches 0, there is a decline in the number of multisensory neurons. When the ρ is 0.1, the number of multisensory neurons is almost zero, leading to a significant reduction in cross-sensory interaction. The weighted accuracy under such conditions is 89.89% and 91.24%, which is also lower than that in other conditions.

	ρ	nom	p	Acc
RAVDESS	0.1	1	2%	0.8989
	0.3	23	6.1%	0.9074
	0.5	55	15.9%	0.9596
	0.7	99	32.9%	**0.9963**
	0.9	163	68.8%	0.9426
eNTERFACE'05	0.1	0	0	0.9124
	0.3	16	4.2%	0.9224
	0.5	52	14.9%	0.9380
	0.7	103	34.7%	**0.9767**
	0.9	164	69.5%	0.9147

Table 3: Comparison of accuracy for neuronal diversity on datasets. nom: Number of multisensory neurons. p: the proportion of multisensory neurons among all neurons. Acc: the weighted accuracy.

Moreover, with ρ set to 0.9, the number of multisensory neurons constitutes over 60% of the total number of neurons. The performance under such conditions is inferior to situations where the number of multisensory neurons accounts for approximately 33%. This is due to an imbalance between the number of unisensory neurons for each sense and multisensory neurons. This imbalance arises from an overemphasis on the interaction of features from one modality on the other, neglecting the extraction of features from the other modality. Therefore, our model performs optimally when ρ is set to 0.7, achieving a relative equilibrium in the number of different types of neurons.

Conclusion

In this study, we introduce a novel multisensory recognition model (ND-MRM), drawing inspiration from diverse neurons that exhibit distinct responses to sensory inputs. Mirroring the neuronal diversity observed in the human brain, our model incorporates both unisensory neurons and multisensory neurons. This unique configuration enables the ND-MRM to effectively model and learn cross-sensory interaction, facilitating multisensory recognition. Additionally, special connection constraints are devised to govern feature transmission, capturing the diverse response characteristics of neurons.

The ND-MRM is assessed in the context of multisensory emotion recognition tasks, and experimental results underscore its superiority over other brain-inspired approaches. However, it's worth noting that our study focuses solely on the auditory and visual modalities of the emotion recognition task. For future research, we plan to extend our method to incorporate multiple other modalities, exploring its effectiveness and adaptability across diverse tasks. Additionally, the consideration of an end-to-end model remains a prospect for future research.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62236001 and Grant 62325601.

References

Allman, B. L.; Keniston, L. P.; and Meredith, M. A. 2009. Not just for bimodal neurons anymore: the contribution of unimodal neurons to cortical multisensory processing. *Brain topography*, 21: 157–167.

Alvarado, J. C.; Vaughan, J. W.; Stanford, T. R.; and Stein, B. E. 2007. Multisensory versus unisensory integration: contrasting modes in the superior colliculus. *Journal of neurophysiology*, 97(5): 3193–3205.

Benssassi, E. M.; and Ye, J. 2023. Investigating Multisensory Integration in Emotion Recognition Through Bio-Inspired Computational Models.

Bhattacharjee, D.; Zhang, T.; Süsstrunk, S.; and Salzmann, M. 2022. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12031–12041. New Orleans, LA, USA: IEEE.

Boehm, K. M.; Aherne, E. A.; Ellenson, L.; Nikolovski, I.; Alghamdi, M.; Vázquez-García, I.; Zamarin, D.; Long Roche, K.; Liu, Y.; Patel, D.; et al. 2022a. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature cancer*, 3(6): 723–733.

Boehm, K. M.; Khosravi, P.; Vanguri, R.; Gao, J.; and Shah, S. P. 2022b. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2): 114–126.

Chabrol, F. P.; Arenz, A.; Wiechert, M. T.; Margrie, T. W.; and Digregorio, D. A. 2015. Synaptic diversity enables temporal coding of coincident multisensory inputs in single neurons. *Nature Neuroscience*, 18(5): 718.

Chumachenko, K.; Iosifidis, A.; and Gabbouj, M. 2022. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2822–2828. IEEE.

Cuppini, C.; Magosso, E.; and Ursino, M. 2011. Organization, maturation, and plasticity of multisensory integration: insights from computational modeling studies. *Frontiers in psychology*, 2: 77.

Cuppini, C.; Stein, B. E.; Rowland, B. A.; Magosso, E.; and Ursino, M. 2011. A computational study of multisensory maturation in the superior colliculus (SC). *Experimental Brain Research*, 213(2-3): 341–349.

Gong, W.; Wang, Y.; Zhang, M.; Mihankhah, E.; Chen, H.; and Wang, D. 2021. A fast anomaly diagnosis approach based on modified CNN and multisensor data fusion. *IEEE Transactions on Industrial Electronics*, 69(12): 13636–13646.

Gültekin, Ö.; Cinar, E.; Özkan, K.; and Yazıcı, A. 2022. Multisensory data fusion-based deep learning approach for fault diagnosis of an industrial autonomous transfer vehicle. *Expert Systems with Applications*, 200: 117055.

Heredia, J.; Lopes-Silva, E.; Cardinale, Y.; Diaz-Amado, J.; Dongo, I.; Graterol, W.; and Aguilera, A. 2022. Adaptive multimodal emotion detection architecture for social robots. *IEEE Access*, 10: 20727–20744.

Holmes, N. P. 2007. The law of inverse effectiveness in neurons and behaviour: multisensory integration versus normal variability. *Neuropsychologia*, 45(14): 3340–3345.

Jessen, S.; and Kotz, S. A. 2013. On the role of crossmodal prediction in audiovisual emotion perception. *Frontiers in Human Neuroscience*, 7: 369.

Jia, S.; Zhang, T.; Cheng, X.; Liu, H.; and Xu, B. 2021. Neuronal-plasticity and reward-propagation improved recurrent spiking neural networks. *Frontiers in Neuroscience*, 15: 654786.

Jia, S.; Zuo, R.; Zhang, T.; Liu, H.; and Xu, B. 2022. Motif-Topology and Reward-Learning Improved Spiking Neural Network for Efficient Multi-Sensory Integration. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8917–8921. Singapore, Singapore: IEEE.

Ju, X.; Zhang, D.; Li, J.; and Zhou, G. 2020. Transformer-Based Label Set Generation for Multi-Modal Multi-Label Emotion Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 512–520. New York, NY, USA: Association for Computing Machinery.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32.

Laurienti, P. J.; Perrault, T. J.; Stanford, T. R.; Wallace, M. T.; and Stein, B. E. 2005. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166: 289–297.

Lee, M.; and Pavlovic, V. 2021. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 1692–1700. Nashville, TN, USA: IEEE.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5): 1–35.

Mansouri-Benssassi, E.; and Ye, J. 2020. Synch-Graph: Multisensory Emotion Recognition Through Neural Synchrony via Graph Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02): 1351–1358.

Martin, O.; Kotsia, I.; Macq, B.; and Pitas, I. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. In *22nd*

International Conference on Data Engineering Workshops (ICDEW'06), 8–8. Atlanta, GA, USA: IEEE.

McDonald, J. J.; Teder-Salejarvi, W. A.; and Ward, L. M. 2001. Multisensory integration and crossmodal attention effects in the human brain. *Science*, 292(5523): 1791–1791.

Molholm, S.; Ritter, W.; Murray, M. M.; Javitt, D. C.; Schroeder, C. E.; and Foxe, J. J. 2002. Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive brain research*, 14(1): 115–128.

Murray, M. M.; and Wallace, M. T. 2011. *The neural bases of multisensory processes*. CRC Press.

Ohshiro, T.; Angelaki, D. E.; and DeAngelis, G. C. 2011. A normalization model of multisensory integration. *Nature neuroscience*, 14(6): 775–782.

Papanastasiou, S.; Kousi, N.; Karagiannis, P.; Gkournelos, C.; Papavasileiou, A.; Dimoulas, K.; Baris, K.; Koukas, S.; Michalos, G.; and Makris, S. 2019. Towards seamless human robot collaboration: integrating multimodal interaction. *The International Journal of Advanced Manufacturing Technology*, 105: 3881–3897.

Shin, A.; Ishii, M.; and Narihira, T. 2022. Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision. *International journal of computer vision*, 130(2): 435–454.

Stein, B. E. 2012. *The new handbook of multisensory processing*. Mit Press.

Stein, B. E.; and Meredith, M. A. 1993. *The merging of the senses*. MIT press.

Stein, B. E.; and Stanford, T. R. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(5): 255–266.

Stein, B. E.; Stanford, T. R.; and Rowland, B. A. 2014. Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience*, 15(8): 520–535.

Stevenson, R. A.; Ghose, D.; Fister, J. K.; Sarko, D. K.; Altieri, N. A.; Nidiffer, A. R.; Kurela, L. R.; Siemann, J. K.; James, T. W.; and Wallace, M. T. 2014. Identifying and quantifying multisensory integration: a tutorial review. *Brain topography*, 27(6): 707—730.

Suzuki, M.; and Matsuo, Y. 2022. A survey of multimodal deep generative models. *Advanced Robotics*, 36(5-6): 261–278.

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, 6558. NIH Public Access.

Wang, Y.; Wu, J.; Furumai, K.; Wada, S.; and Kurihara, S. 2022. VAE-based adversarial multimodal domain transfer for video-level sentiment analysis. *IEEE Access*, 10: 51315–51324.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017a. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915. Venice, Italy: IEEE.

Zhang, M.; Li, W.; Tao, R.; Li, H.; and Du, Q. 2021. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; and Tian, Q. 2017b. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 3030–3043.

Zhang, Y.; Chen, M.; Shen, J.; and Wang, C. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9100–9108. AAAI Press.

Zhu, Z.; Li, Y.; Lyu, W.; Singh, K. K.; Shu, Z.; Pirk, S.; and Hoiem, D. 2023. Consistent Multimodal Generation via A Unified GAN Framework. arXiv:2307.01425.