Dean&Francis
ISSN 2959-6157

Research Progress and Application Prospects of Multimodal Brain-Computer Interface Technology

Xin Hong

Abstract:

In the past decades, the rapid development of brain-computer interface (BCI) technology has provided new perspectives on human-computer interaction. Traditional single-modal BCI systems, although showing promising results in many applications, still face challenges in terms of decoding accuracy, real-time performance, and user adaptability due to limitations in signal acquisition. To address these issues, multimodal brain-computer interfaces (MMBCIs) have emerged, aiming to improve the overall system performance by integrating information from different physiological signal sources. This paper reviews the basic concepts, signal sources, signal fusion techniques, application examples, and current challenges and future perspectives of multimodal BCI. By analyzing the existing research literature, we hope to provide theoretical guidance and framework support for further research in this area.

Keywords: brain-computer interface, multimodal signal fusion, deep learning, human-computer interaction, signal decoding, neuroscience, data processing techniques

1. Introduction

With the rapid development of neuroscience and information technology, Brain-Computer Interface (BCI) technology has gradually become an important research direction in the field of human-computer interaction. BCI technology greatly expands the way human beings connect with the digital world by directly decoding the user's brain signals, thus allowing barrier-free interaction between the brain and external devices. Although single-modal BCI systems have demonstrated some practical value in specific application scenarios, such as decoding motor intention using electroencephalogram (EEG) signals or monitoring changes in cerebral blood flow using functional near-infrared spectroscopy (fNIRS), such systems are often limited due to a variety of factors, mainly in terms of signal accuracy, real-time performance, and user adaptability.

Multimodal BCI is proposed to address these limitations. The multimodal BCI system integrates information from different physiological signal sources, such as EEG and fNIRS, to complement each other's strengths and weaknesses, with the expectation of improving the accuracy of signal decoding and the stability of the system. For example, the advantage

ISSN 2959-6157

of EEG signals in temporal resolution can compensate for the deficiency of fNIRS in temporal delay, and vice versa; such signal fusion can form a richer user brain state substrate, which improves the stability and accuracy of the multimodal BCI system in complex interference environments.

In recent years, with the rapid progress of multiple data processing technologies such as deep learning, how to effectively fuse information from multiple signal sources has become a hot spot in research. The research progress and application prospects of multimodal brain-computer interface technology mainly focus on the fusion and conversion of multimodal information through deep learning to improve model performance. At the early stage of development, the multimodal fusion technology focuses on improving the classification and regression ability of deep learning models, and the application of architectures such as joint, collaborative, and decoder, as well as their advantages and disadvantages, are analyzed by elaborating the multimodal fusion architectures, methods, and alignment techniques. Specific methods include multi-kernel learning, image modeling, and neural networks, etc., which utilize public datasets for cross-modal transfer learning and modal semantic conflict resolution. It is shown that the performance of the model can be effectively improved by comprehensively utilizing multi-domain information such as text, image, speech, and video. The core of multimodal technology lies in representation, fusion, transformation, and alignment, which can better utilize multimodal data in a comprehensive way by projecting heterogeneous features into a common subspace. Future research directions include further optimizing cross-modal learning, resolving multimodal semantic conflicts, and improving multimodal combination evaluation capabilities.

The purpose of this study is to systematically review the latest research progress and application examples of multimodal BCI technology and to look forward to future research directions. This research will not only make a substantial contribution to promoting the development of BCI technology but will also provide theoretical and practical support for the fields of neurorehabilitation, affective computing, and intelligent human-computer interaction.

2. Sources of information

2.1 Visual information processing

Vision is one of the most important ways for people to perceive the external world. Visual signals are received through the retina and then transmitted to the cerebral cortex for processing, finally forming the picture that people are aware of. Visual information processing is an important direction in the research of multimodal brain-computer interface technology. Through the measurement of human eye electrical signals, human eye movement, pupil size, and other information are transformed into signals that can be understood by computers, thus allowing visual information to be presented and processed in a digital way. Digital vision, on the other hand, is an important perceptual medium in this interactive process. In the emerging fields of virtual reality (VR), augmented reality (AR), mixed reality (MR), and so on, the integration of digital vision and multimedia interaction technology is particularly significant. Through high-precision 3D modeling, real-time rendering, and somatosensory interaction technologies, users can explore and manipulate the virtual world in an immersive manner, realizing a seamless information interaction experience with the real environment. In addition, in 1978, William Dobelle, a trailblazer in the field of visual brain-computer interfaces, introduced an innovative technique by placing a grid of 68 electrodes into the visual cortex of Jerry, who was blind. This groundbreaking procedure resulted in the generation of phosphene, a form of visual perception. The brain-computer interface system is composed of a video camera for capturing images, a signal processing unit, and electrodes for cortical stimulation. Once implanted, the patient is able to perceive low-resolution images with a dot-matrix format and grayscale modulation, although the visual field is restricted and the refresh rate is low. This visual prosthesis system is designed to be portable, allowing patients to use it autonomously without needing help from a physician or technician.

2.1.1 Eye movements serve

Eye movements serve as a direct reflection of our thoughts, objectives, and memories, significantly influencing our interpretation of the visual environment. Consequently, integrating eye tracking with neuroimaging techniques offers valuable insights into various aspects of human cognition, including conditions like neurodegenerative diseases and neurological disorders. Functional Magnetic Resonance Imaging (fMRI) is a prominent method for investigating brain function in humans, enabling researchers to analyze brain activity while participants engage in different tasks. In numerous fMRI studies, the observed behaviors can either be a focal point of interest or a possible confounding factor. Nevertheless, eye tracking is often neglected in most of these studies.

Looking back 30 years, the history of eye-tracking technology can be roughly divided into three phases:

2000 years ago, the study of the human eye's point of gaze, which began as early as the 19th century, was primarily applied in physiology, psychology, and related

academic fields of ophthalmology for the purpose of understanding how the human eye works and how people process information both consciously and unconsciously (Javal, 1990).

2000~2020, this stage with the rise of the IT industry and so on, the Internet economy is almost equivalent to the attention economy, also known as the eyeball economy, accompanied by the miniaturization of eye-tracking technology, lightweight, more and more used in web applications, advertising and marketing and other fields.

After 2020, eye-tracking technology and other application areas become more extensive, especially the near-eye display form of XR devices began to integrate eye-tracking technology, the most representative of the AR glasses HoloLens 2 from Microsoft and widely used in scientific research HTC VIVE Pro Eye, both released in 2019.

2.1.2 techniques and data-base

There are a number of techniques for implementing eye tracking, including but not limited to:

1. Electrooculogram (EOG)
2. Scleral Electromagnetic Tracking Coil
3. Video based pupil monitoring
4. Infrared Corneal Reflection

XR Near-eye display devices basically use the infrared corneal reflection method, which is simply to utilize the difference between the cornea and the iris in reflecting near-infrared light and to capture and calculate the direction of eye movements by using a near-infrared fill light and near-infrared camera (Yan Guoli, Bai Xuejun, 2018). The following is the eye movement tracking dataset:

Table 1

Data	Sub.	Tracker			FRQ	Res.	Num.	Seg 2D			Seg 3D			PC	LM 2D			LM 3D			Eye	Ga	Mov.			
		VR	AR	HM			Annot	P	I	Sc	P	I	Sc		P	I	Lid	P	I	Lid			F	S	SP	B
POG[75]	20	-	-	1	30Hz	768 × 480	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-
NNVEC[30]	20	-	-	1	25Hz	384 × 288	866,069	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-	-	-
NVGaze[70]	35	1	1	-	120Hz	640 × 480	2,500,000	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-	-	-
Casia.v1[77, 92]	108	-	-	1	-	320 × 280	756	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casia.v2[77, 92]	60	-	-	2	-	640 × 480	2,400	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casia.v3[77, 92]	≈700	-	-	3	-	Multiple	22,034	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casia.v4[77, 92]	≈1,800	-	-	4	-	Multiple	54,601	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casia.test[77, 92]	1,000	-	-	1	-	640 × 480	10,000	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Casia.age[3, 94]	50	-	-	2	-	Multiple	≈160,000	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ubiris.v1[78]	241	-	-	Y	30Hz	Multiple	877	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ubiris.v2[79]	261	-	-	Y	200Hz	400 × 300	≈11,000	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MASD[10–12]	82	-	-	Y	-	Multiple	2,624	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GAN[32]	22	-	-	1	120Hz	640 × 480	130,856	Y	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
500k[50]	20	-	-	1	25Hz	384 × 288	866,069	Y	-	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MEMD[41]	20	-	-	1	25Hz	384 × 288	866,069	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	-
ME[96]	587	1	1	-	-	640 × 480	880,000	Y	Y	Y	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-
OpenEDS[63]	152	1	-	-	200Hz	400 × 640	356,649	Y	Y	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GIW[72]	19	-	-	1	120Hz	640 × 480	≈2,016,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y
BAY[86]	6	-	-	1	30Hz	640 × 480	27,022	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	-
HEV[13]	57	1	-	-	24-30Hz	-	no images	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-
HEI[82]	63	1	-	-	60Hz	-	no images	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-
LPW[93]	22	-	-	1	120Hz	640 × 480	130,856	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-
Swi[89]	2	-	-	1	-	620 × 460	600	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-
ExCuSe[46]	7	-	-	1	25Hz	384 × 288	39,001	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-
Else[58]	17	-	-	1	25Hz	384 × 288	55,712	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-
PNET[56, 57]	5	-	-	1	25Hz	384 × 288	41,217	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-	-	-	-
EWO[53]	11	-	-	1	25Hz	384 × 288	1,100	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-
FRE[54]	11	-	-	1	25Hz	384 × 288	4,000	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	-
TEyeD	39	-	-	1	25Hz	384 × 288	5,665,053	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	1	-	-	1	60Hz	320 × 240	12,184	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	22	-	-	1	95Hz	640 × 480	130,856	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	54	1	1	1	120Hz	640 × 480	8,691,764	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	16	-	-	1	60Hz	640 × 360	6,367,216	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

2.1.3 Classification of eye-tracking interaction applications.

Today there are dialing/unlocking interactions using eye movements. For example, apple's ios18 system, Apple Vision Pro Eye-Hand Synergy is also an active interaction based on eye tracking that

In addition to this interface input interaction, eye movement input can also be used for game controls, such as weapon switching in PSVR 2 games

2.1.3 .1. Passive

ISSN 2959-6157

a technique that optimizes the rendering of a picture by tracking the position of the eye's gaze in real-time. For example, gaze point rendering, only in the human eye's most visually sensitive central concave (Foveal) region to render the highest resolution, with the distance away from the central concave increased visual sensitivity will also be sharply reduced, and accordingly only render a lower resolution of the screen, thus greatly reducing the burden of the headset screen rendering. b Based on the gaze point to achieve the auto-zoom function, including the Apple Vision Pro, Currently, all known head-up display screens have fixed focal length (usually 1 ~ 1.5m), the screen light has no depth information, the convergence and focusing position of the separation occurred, thus generating visual convergence adjustment conflict (VAC problem), triggering visual fatigue, dizziness, and other problems. The focus point zoom can dynamically adjust the optical focus according to the content of the user's attention, thus realizing a more comfortable and natural visual experience.

2.1.3 .2. Expressive & Diagnostic Interaction

Apple Vision Pro's Eyesight feature is also an expressive application based on eye tracking, which uses an internal camera to track the user's real-time eye movements before re-modeling them and rendering them on an external screen to reduce the barrier between the wearer and the person next to them.

2.1.4 Performance Requirements

Spatial Resolution and Temporal Resolution are two dimensions that disentangle the eye-tracking performance requirements of different application scenarios, where Spatial Resolution includes Accuracy and Precision; and Temporal Resolution includes Sampling Rate and End-to-End Latency. Spatial resolution includes Accuracy and Precision, while temporal resolution includes Sampling Rate and End-to-End Latency.

2.2 Auditory information processing

Auditory information processing is also an important direction in the research of multimodal brain-computer interface technology. By measuring human auditory responses, such as brain waves, gibbous evoked potentials, and other information, auditory information is converted into signals that can be understood by computers in order to realize speech recognition, sound synthesis, and other functions. For example, this technology can be used to realize the voice control function of the human-computer interaction interface.

2.2.1 speech recognition

Speech is an analog signal, which needs to be sampled and processed by the microphone (array) and other equipment to become a digital signal that can be processed by the machine; then after feature extraction, the signal is converted from the time domain to the frequency domain; and then using the extracted feature vectors, it is finally converted into text after pattern matching. Among them, the acoustic model and language model of the pattern matching link determine the final recognition effect, the acoustic model and language model need to be trained using the labeled data, and currently most of the supervised learning algorithms are used to achieve the advantage of high accuracy, the disadvantage is the need for human intervention and large workload.

In practical applications, in addition to focusing on the machine learning algorithms and software processing used, it is also necessary to pay attention to the speech acquisition link, especially in far-field interaction scenarios (e.g., smart audio). Speech acquisition is the precondition for speech recognition, if the quality of the acquired speech is not high, even if the arithmetic power is sufficient, the algorithm is subtle, and the data volume is large and high quality, the final recognition accuracy is not ideal. Therefore, in some scenarios, it is necessary to improve the quality of voice acquisition by improving the microphone noise reduction effect and adopting microphone arrays.

The application of deep learning algorithms has made speech recognition one of the first breakthroughs in artificial intelligence. Currently, the speech recognition accuracy of mainstream market vendors exceeds the human level, with the optimized recognition accuracy of cell phones, computers, noise-canceling microphones (arrays), and other scenarios being higher than 95%, the optimized recognition accuracy of phone calls and other scenarios being higher than 85%, and the optimized recognition accuracy of other voices (compressed stored recordings and other scenarios) being higher than 80%.

2.2.2 Data sets:

1. International data set
1.1 .LJSpeech (single)
1.2 .JSUT
1.3 .RUSLAN
1.4 .RyanSpeech
1.5 .VocBench
1.6 Arabic Speech Corpus
1.7 .Silent Speech EMG
1.8 .Hi-Fi Multi-Speaker English TTS Dataset
1.9 .KSS (single)
1.1 0.VCTK (multiple)
1.1 1.LibriTTS(multiple)
2. Multitask data set
2.1 .ESD dataset

3. Speech Sentiment Recognition Data set
3.1 .IEMOCAP (The Interactive Emotional Dyadic Motion Capture Database)
2.2.3 Natural Language Generation & Language Synthesis
Natural Language Processing (NLP) is an important branch of Artificial Intelligence that focuses on computers understanding and generating human language. Multimodal learning, on the other hand, refers to the ability of a machine learning model to process different types of data, such as images, audio, text, and so on. With the development of deep learning technology, multimodal learning has received widespread attention in the field of NLP.
The integration of multimodal learning with NLP can help machines better understand human language and produce better results in many application scenarios. For example, in the image description task, the model needs to understand the content in the image and generate the corresponding text description; in the sentiment analysis task, the model needs to recognize the sentiment tendency from the text and classify it; in the machine translation task, the model needs to understand the semantic relationship between two different languages.
A speech neuroprosthesis is a device designed to interpret brain activity related to speech and convert it into communicative outputs, which can include text, sounds, or facial movements associated with speech. These neuroprostheses not only facilitate more natural forms of communication but also help restore other expressive elements that convey meaning, such as tone, volume, and facial expressions. Recent advancements in speech neuroscience, neural interface technology, and machine learning have propelled the development of clinically practical speech neuroprostheses. Ongoing research has deepened the understanding of how speech features are represented in the cortex, with a particular focus on the motor control mechanisms involved in the vocal tract, which can assist patients with incomplete atresia syndrome who rely on Augmentative and alternative communication (AAC , AAC) methods of communication, by laying the groundwork for the decoding of their cortical activity into textual Laying the groundwork for the development of a verbal neuroprosthesis.

2.3 Haptic information processing

Compared with visual and auditory information, haptic information processing has been studied relatively little. However, with the application of more and more wearable technologies, the study of human tactile responses has triggered attention to haptic information processing in multimodal brain-computer interfaces. Currently, researchers have realized a certain degree of haptic feedback through technologies such as skin sensors and artificial arms.
Since their discovery in 2010, Piezo proteins have been shown to be widely distributed in various tissues and organs of the human body. Ion channels composed of Piezo proteins as subunits respond to specific mechanical stimuli and open, allowing positively charged ions to flow into the cell. The two ion channels in the Piezo protein family, Piezo1 and Piezo2, have different activation mechanisms. Piezo1 can be activated by either positive or negative pressure, while Piezo2 can only be activated by positive pressure. Their distribution in the body also differs; Piezo1 is primarily found in non-sensory tissues, particularly in fluid dynamic pressure environments, providing mechanosensitivity to non-excitable cells, participating in erythrocyte volume regulation, shear stress sensing, and perceiving fluid flow in the kidneys. Conversely, Piezo2 is mainly expressed in sensory nerve cells and is involved in mechanosensitivity related to touch and proprioception, as well as in the mechanosensory function of neurons and their environments. Both Piezo1 and Piezo2 also participate in stress perception at specific sites, such as articular chondrocytes. Researchers have developed several techniques to stimulate Piezo protein ion channels in vitro, including “stretching” and “poking” based on membrane-clamp electrophysiology, indentation and shear force testing using atomic force microscopy (AFM), and approaches involving chemical agonists or magnetic nanoparticles. However, due to the advantages and disadvantages of each technique in terms of ease of use, the number of sampled channels, and the quantification of stimuli and responses, the exact mechanism of how mechanical forces couple to the activation of ion channels has yet to be elucidated. The fusion of touch in multimodal learning has been a challenge due to the expensive process of tactile data collection and the poor standardization of sensor outputs. A recent paper, Binding Touch to Everything: Learning Unified Multimodal Tactile Representations, presented by a team of researchers from Yale University and the University of Michigan, offers a new solution to this challenge. The team proposed a unified haptic model called UniTouch that connects vision-based tactile sensors to multiple modalities.UniTouch does this by aligning its embeddings with pre-trained image embeddings already associated with other modalities.

2.3.1 Zero-sample haptic understanding

Urgent alignment of haptics and text enables zero-sample haptic understanding, e.g., material categorization and grasp stability prediction. Following CLIP, the authors encoded haptic images and textual cues using templates and class names. Zero-sample classification is achieved by

ISSN 2959-6157

computing and ranking similarity scores between them.

2.3.2 Haptic-LLM (Language Model)

using an existing visual-linguistic model that is consistent with the image embedding of the aligned haptic embedding, a haptic-linguistic model can be created by switching to the present haptic encoder. Given haptic images and linguistic inputs, a fuller understanding can be obtained through questions and answers.

2.3.3 Haptic Image Synthesis

Binding haptics to text also opens up additional potential capabilities for haptic image synthesis. The authors utilize a pre-trained text-to-image diffusion model and use haptic features to condition the denoising process for zero-sample haptic-to-image generation [and haptic-driven image stylization.

2.3.4 Database

Table 2

	Dataset	Sensor	# data	Material cls.	Robot grasp
Train & Eval	Touch and Go [111]	GelSight	120k	✓	
	The Feeling of Success [6]	GelSight	9.3k		✓
	YCB-Slide [94]	DIGIT	183k	✓	
	Object Folder 2.0 [32]	Taxim	180k	✓	✓
Eval.	Object Folder Real [33]	GelSlim	20k	✓	
	Object Folder 1.0 [30]	TACTO	20k	✓	✓
	SSVTP [57]	DIGIT	4.6k	✓	

Table 1. **Datasets for training and evaluation.**

3. Brain-computer interface branches

Multimodal neural interfaces primarily involve the use of optical, electrical, magnetic, acoustic, and chemical drug delivery methods to record and modulate neural activity.

3.1 Optical Neural Interface

Among them, optical neural interfaces mainly include optical recording of neural activity and optical stimulation, which are categorized into exogenous and endogenous modalities according to whether exogenous materials (including photosensitive genes, nanomaterials, dye compounds, etc.) are applied.

The main principle of exogenous optical recording is to correlate fluorescent signals with neuronal local potentials by molecular synthesis or genetic engineering and to observe and record fluorescent signals from specific "active" wavelengths by optical detection. Fluorescent probes are widely used in neuroscience and generally include synthetic, genetically encoded, and hybrid (using a combination of synthetic dyes and genetically encoded proteins). Now, in combination with large-scale single-photon or multiphoton imaging, we are able to read the activity of neural circuits during wakefulness and in correlation with animal behavioral science. Exogenous optical stimuli, on the other hand, include photoactive molecules based on photochemical interactions, nanomaterials based on photothermal and photovoltaic interactions, and optogenetic stimuli that require genetic modification; whereas endogenous optical signal (OS) recordings comprise both direct measurements of the scattering properties of light interacting with neural tissues, and indirect measurements of changes in the concentration of labeled substances that correlate with brain activity. Endogenous optical stimulation mainly utilizes the sensitivity properties of certain neurons themselves to specific light conditions for optical modulation and so on.

3.2 Electrical Neural Interface

Electrical neural interfaces serve as direct communication pathways between the nervous system and external devices. Recent technological developments in this area have led to more effective tools for studying, restoring, and enhancing neural functions. Nonetheless, the complex structure of the nervous system presents considerable challenges in the design, fabrication, and integration of these systems. This review emphasizes recent progress in neuroelectrical interfaces, particularly focusing on innovative technologies that enhance both spatiotemporal resolution and the ability to map and manipulate brain circuits. Key topics include large-scale, long-term neural recording, wireless, and miniaturized implantation techniques, as well as advancements in signal transmission, amplification, and processing, alongside the integration of electrical interfaces with optical technologies.

Neural interface technology using electricity has become

critical in basic neuroscience and translational medicine. Common neurophysiological signals utilized include electroencephalograms (EEG), electrocorticograms (ECoG), local field potentials (LFPs), action potentials (APs), and spike potentials, which are primarily obtained through microelectrodes. Implantable microelectrodes and associated technologies are preferred for contemporary neural interface applications due to their sub-millisecond temporal resolution and capability to detect individual neuronal electrical signals in vivo.

Notably, flexible microelectrodes that have surfaced in recent years offer excellent nerve-electrode interfaces, facilitating long-term stable large-scale neural recordings. This paper examines the latest advancements in both optical and electrical neural interfaces, covering the principles of optical recording and stimulation through genetic engineering techniques, hemodynamic imaging, as well as the recording and stimulation capabilities of implantable microelectrodes, among other methodologies.

3.3 Magnetic neural interfaces

Magnetic neuromodulation provides a wireless, remote means of deep brain stimulation, offering advantages over optogenetic and wired electrode methods. However, its adoption has been limited due to a lack of understanding regarding its mechanisms and the shortcomings of early magnetic systems. Additionally, while magnetic neuromodulation holds significant promise for neuroscience research, the exploration of cell-type-specific magnetic modulation is still in its infancy.

Researchers at the Institute for Basic Science (IBS) and the Center for Nanomedicine at Yonsei University have developed a novel magnetogenetic technology known as the “Neurokinetic Magnetobiological Interface” (Nano-MIND). This approach employs a nanomaterial-based toolkit that, in conjunction with Cre-loxP technology, selectively activates genetically encoded Piezo1 ion channels in targeted neuronal populations. The technique also utilizes the torque generated by nanomagnetic actuators for neuromodulation both in vitro and in vivo.

The application of this cell type-specific magnetic method has been demonstrated in various behavioral models, including bidirectional regulation of feeding behavior, long-term weight control in obese mice, and wireless modulation of social behaviors among multiple mice in a shared environment. This capability facilitates the remote control of specific brain regions, allowing for the modulation of complex brain functions related to emotions, social interactions, and motivation in animal subjects.

In their studies, the team applied Nano-MIND technology to selectively activate inhibitory GABA receptors in the medial preoptic area (MPOA) of infertile female mice, resulting in a significant enhancement of parental behaviors. Additionally, the technique was used to stimulate motivational circuits in the lateral hypothalamus, affecting the feeding behavior of the animals. The findings revealed that activation of inhibitory neurons in these areas led to a 100% increase in appetite and feeding behavior, while excitation of neurons resulted in a more than 50% reduction in appetite.

These results demonstrate that Nano-MIND technology can selectively engage specific brain circuits and bidirectionally regulate higher brain functions. The researchers anticipate that this technology will advance the understanding of brain function, aid in the development of complex artificial neural networks and bidirectional brain-computer interface systems, and open new pathways for treating neurological disorders.

3.4 Acoustic Neural Interface

A novel noninvasive closed-loop acoustic brain-computer interface (BCI) has been developed to decode the onset time of seizures using EEG signals and to trigger vagal ultrasound stimulation to halt seizures. This recent study established the BCI system and utilized a multilevel threshold model to decode seizure onset from wirelessly collected electroencephalography (EEG) data recorded from the hippocampus.

In an epileptic rat model, the vagus nerve was stimulated using acoustic radiation force, applied via varying acoustic parameters, to evoke responses when pen tetrazole was administered. Subsequently, the EEG signals indicative of seizures initiated ultrasonic stimulation of the vagus nerve to influence the outcomes of the seizures. Additionally, the mechanisms underlying seizure control by the BCI system were examined through real-time quantitative polymerase chain reaction (RT-qPCR) techniques.

4. Existing convergence technologies

4.1 mVEP and MI

The hybrid brain-computer interface (BCI) system successfully generated both expected motor imagery (MI) and modified visual evoked potential (mVEP) signal characteristics, with both signals resembling those produced in a unimodal BCI task. Furthermore, results from online 2D motion control experiments indicate that this hybrid BCI offers more efficient and intuitive control commands.

The significance of this is that the hybrid BCI system introduces a compensation mechanism for effective 2D motion control, particularly in application scenarios where

ISSN 2959-6157

P300 stimuli may not be applicable. Overall, this system enhances the feasibility of online control and opens up new opportunities for BCI applications in complex environments, facilitating smoother and more precise motion control in practical settings.

4.2 EEG and fNIRS

EEG measures the electrical activity of neurons in the cerebral cortex, boasting high temporal resolution that captures rapid changes in neural activity at the millisecond level. This characteristic makes EEG particularly effective for detecting fast brain signals, such as those associated with motor imagery (MI) and event-related potentials (ERP). In contrast, fNIRS assesses neural activity indirectly by monitoring variations in blood oxygen concentration within the brain, offering a high spatial resolution that provides detailed information about the localization of brain regions. It primarily detects the hemodynamic response related to increased neural activity and is effective in capturing slower physiological changes.

While EEG excels in temporal resolution, it has limitations in spatial resolution, making it challenging to precisely locate brain activity. Conversely, fNIRS, with its high spatial resolution, is versatile in its applications but suffers from low temporal resolution, hindering its ability to track rapid fluctuations in brain activity in real-time. By integrating both EEG and fNIRS, a hybrid BCI system can leverage the strengths of each approach to achieve data that possesses both high temporal and spatial resolution, resulting in a more comprehensive understanding of brain activity.

4.3 adaptive

Multi-feature adaptive fusion framework for kernel tracking. A multi-feature description of the target is constructed using a set of sub-models of the target features, and multiple features of the target are integrated into the kernel tracking method by a linear weighting method. According to the similarity between each feature submodel and the current target and background, there is an adaptive weight updating mechanism based on the Fisher divisibility metric. Meanwhile, in order to overcome the drift during the model updating process, a selective updating strategy is proposed based on the divisibility of submodels, which is capable of automatically adjusting the weights of the features in the tracking according to the actual scene changes to realize the adaptive response to the scene changes. The selective submodel updating strategy realizes the adaptation to the changes of the target itself and reduces the influence of the model drift; in addition, this method has the advantages of simplicity and speed. In the experiments, color and LBP texture are selected as the target features, and through the experiments on several real scenes, it is verified that the proposed method is adaptive to the changes of the scene and the target, and is able to achieve robust real-time target tracking.

5. Summarize

Multimodal learning can be understood as mining and analyzing heterogeneous data from multiple sources, for which different models need to be feature extracted and fused in different ways to accurately capture the deep concepts, contexts, and correlations expressed by these modalities.

In this thesis, we systematically review the research progress and application prospects of multimodal brain-computer interface technology. First, the characteristics and applications of information sources such as visual, auditory, and tactile are analyzed through examples, revealing the importance and unique functions of different perceptual modalities in brain-computer interfaces. The combination of these information sources not only enhances the perceptual ability of the system but also dramatically improves the user's interaction experience.

Second, this paper discusses the neural interface branch of brain-computer interfaces and analyzes the latest development of different neural signal acquisition and processing technologies. These techniques have made significant progress in improving signal quality and reducing interference, laying an important foundation for realizing efficient and accurate brain-computer interaction.

Finally, we analyze the existing fusion technologies and point out the potential of multimodal information fusion in the development of brain-computer interfaces. By integrating various information sources, fusion techniques can not only improve the accuracy of signal decoding but also help to build more complex and intelligent application scenarios.

To summarize, multimodal brain-computer interface technology is in a stage of rapid development and has a promising future application in medical, education, entertainment and other fields. With the continuous progress of technology and in-depth research, we expect that this field will bring more innovative solutions and practical applications to further promote the boundaries of human-computer interaction.

References

[1] Adhanom, I. B., MacNeilage, P., & Folmer, E. (2023). Eye Tracking in virtual reality: a broad review of applications and challenges. Virtual Reality, 1-25.

[2] Belardinelli, A. (2023). Gaze-based intention estimation: principles, methodologies, and applications in HRI. arXiv preprint arXiv:2302.04530.
[3] Choi, SH., Shin, J., Park, C. et al. In vivo magnetogenetics for cell-type-specific targeting and modulation of brain circuits. Nat. Nanotechnol. 19 nat. nanotechnol. 19 , 1333-1343 (2024). https://doi.org/10.1038/s41565-024-01694-2
[4] Duchowski, A. T. (2018). Gaze-based interaction: a 30 year retrospective. Computers & Graphics, 73, 59-69.
[5] Davis, J., Hsieh, Y. H., & Lee, H. C. (2015). Humans perceive flicker artifacts at 500 Hz. scientific reports, 5(1), 7861.
[6] Fazli S, Mehnert J, Steinbrink J, et al. Enhanced performance by a hybrid NIRS-EEG brain computer interface[J]. Neuroimage, 2012, 59(1): 519-529.
[7] Lee, W. J., Kim, J. H., Shin, Y. U., Hwang, S., & Lim, H. W. (2019). Differences in eye movement range based on age and gaze direction.Eye, 33(7), 1145-1151.
[8] Ma T, Li H, Deng L, et al. The hybrid BCI system for movement control by combining motor imagery and moving onset visual evoked potential[J]. Journal of neural engineering, 2017, 14(2): 026015
[9] Ma Teng. Research on key technologies of multimodal brain-computer interface based on mVEP and MI [D]. University of Electronic Science and Technology, 2018.
[10] Melissa R. Warden1,2, Jessica A. Cardin5,6, and Karl Deisseroth2,3, Optical Neural Interfaces (2014) ANNUAL REVIEW OF BIOMEDICAL ENGINEERING Vol 16:103-129
[11] Silva, A.B., Littlejohn, K.T., Liu, J.R. et al. The speech neuroprosthesis. Nat. Rev. Neurosci. 25, 473-492 (2024). https://doi.org/10.1038/s41583-024-00819-9
[12] Silva, A.B., Littlejohn, K.T., Liu, J.R. et al. The speech neuroprosthesis. Nat. Rev. Neurosci. 25, 473-492 (2024). https://doi.org/10.1038/s41583-024-00819-9
[13] XIAO X, XIN F, MEI J, et al. A Review of Adaptive Brain-Computer Interface Research[J]. Journal of Electronics and Information, 2023, 45(7): 2386-2394.
[14] Yan HuangJiaxi XuXuehao ZhangEnyan Yu Research progress on vestibular dysfunction and visual-spatial cognition in patients with Alzheimer's disease Article (Apr 2023) 35(19):15-20
[15] Zou J, Chen H, Chen X, Lin Z, Yang Q, Tie C, Wang H, Niu L, Guo Y, Zheng H. Noninvasive closed-loop acoustic brain-computer interface for seizure control Theranostics 2024; 14(15):5965-5981. doi:10.7150/thno.99820. https://www.thno.org/v14p5965.htm
[16] Zhu Guangming. Research on feature extraction and classification method for multimodal brain-machine interface based on EEG and fNIRS[D]. Hangzhou: Hangzhou University of Electronic Science and Technology, 2017.