This article has been accepted for publication in a future proccedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICIT58465.2023.10143111, 2023 IEEE International Conference on Industrial Technology (ICIT)

Circuit Design of Multimodal Attention Memristive Network for Affective Video Content Analysis

Xiaoyue Ji
College of Electrical Engineering
Zhejiang University
Hangzhou, China
ji.xiaoyue@zju.edu.cn

Zhekang Dong*
School of Electronics & Information
Hangzhou Dianzi University
Hangzhou, China
englishp@hdu.edu.cn

Chun Sing Lai
Department of Electronic & Electrical Engineering
Brunel University London
London, UK
chunsing.lai@brunel.ac.uk

***Abstract*—Affective video content analysis aims at automatically identifying human emotion triggered by video, which plays an important role in mental health monitoring. This paper proposes a multimodal attention memristive network for affective video content analysis, which offers an energy-efficient approach with low time consumption and high classification accuracy. To illustrate the complexity of the proposed multimodal attention memristive network, two core modules are proposed. Firstly, unimodal feature representation module with cascaded configuration is designed to capture unique characteristics from multimodal signals. Then, multimodal local-global fusion module is proposed to stimulate the process of multimodal information sensing and processing in human brain. Furthermore, the proposed system is validated by applying it to affective content analysis. The experimental results demonstrate that the multimodal attention memristive network outperforms the existing state-of-the-art methods with high classification accuracy and low time consumption.**

Keywords—Circuit design, memristive network, affective video content analysis

I. Introduction

With the rapid development of artificial intelligence (AI) technology and multimedia technology, videos have been becoming a popular information carrier for communication, entertainment, and instruction [1]. Emotion information in video will inevitably reflect user's feelings and mental health [2]. Thus, affective video content analysis aims at establishing the relationship between multimodal information and affective concepts, which is now attracting increasing attention in recent years [3].

Recent researches of affective video content analysis are mainly based on machine learning methods and deep learning methods [4-8]. An improved method for affective video content analysis based on domain knowledge was proposed in [4], which successfully improved emotion recognition accuracy by using well-established film grammar. A multimodal deep regression Bayesian network was constructed to extract the dependencies between aural elements and visual elements for affective video content analysis [5]. A multimodal learning framework for video content analysis was designed in [6], which classified affective contents in the valence-arousal space. In [7], sentiment driven features was used to classify human emotion states in videos, which outperformed the state-of-the-art methods in terms of valence and arousal classification accuracy. Inspired by the multimodal integration effect, a multimodal local-global attention network was proposed in [8], which taken a novel four modality representation (i.e., visual, audio, motion, and tone) of video as input for affective content analysis. Although, these above-mentioned methods have achieved superior performance in terms of classification accuracy, while have certain limitations in running time and energy consumption.

Memristor is a two-terminal circuit component, exhibiting high density, low power consumption, non-volatility, and synaptic properties, which are potential candidates for establishing ultra-low power consumption neuromorphic computing system for affective video content analysis [9, 10]. A flexible neuromorphic computing system via memristive circuit was built up, which can realize affective communication with high accuracy and low time consumption [11]. A memristor-based hierarchical attention network with low energy consumption, low privacy invasiveness, and low fabrication cost was proposed in [12], which can effectively perform multimodal affective computing in smart home. In [13], multimodal neuromorphic sensory-processing system for indoor human behavior recognition was designed, which offered a more environmentally friendly approach to realize health monitoring in home environment. Inspired by human brain function, an in-memory computing system was developed in [14], which aimed at solving computationally hard problems for von Neumann architectures. A physical-oriented memristor model for bio-inspired computing was constructed, which can realize the automatic conversion from short-term memory (STM) to long-term memory (LTM) [15].

Based on this, this work aims to investigate a multimodal attention memristive network. For verification purposes, the proposed network is applied to affective video content analysis. The main contributions of this work are summarized as follows:

1) Different with existing memristive networks, we present a multimodal attention memristive network that can effectively learn unimodal representations and capture local-global feature from multimodal information.

2) The circuit design of multimodal attention memristive network is proposed using high stability and eco-friendly Ag/TiO_2/FTO memristor, which provides a parallel connection and highly integration to reduce computational cost.

3) The correctness of the proposed multimodal attention memristive network is verified by affective video content analysis. The experimental results demonstrate that the proposed network has good performance in terms of accuracy and time consumption.

The rest of this paper is structured as follows. Section II demonstrates the detailed circuit design of the entire multimodal attention memristive network from the perspectives of unimodal feature representation module and multimodal local-global fusion module. In Section III, the proposed network is applied for affective video content

This article has been accepted for publication in a future proccedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICIT58465.2023.10143111, 2023 IEEE International Conference on Industrial Technology (ICIT)

analysis. Finally, Section IV includes the conclusion drawn from this paper.

II. Circuit Design of Multimodal Attention Memristive Neural Network

A. Overall Circuit Architecture

High accuracy affective video content analysis relies on perceiving and processing multimodal information, including visual, text, and audio [1]. Based on this, we propose a multimodal attention memristive network, which mainly consists of two modules: unimodal feature representation module and multimodal local-global fusion module, as shown in Fig. 1. Specifically, the input multimodal data (i.e., image, text, and audio) are converted into voltage signals by a digital-to-analogue converter (DAC). Then, the corresponding voltage signals need to be injected to the unimodal feature representation module that can fully extract features from the voltage signals (containing image, text and, audio information. To investigate the relationship between unimodal feature representations, the multimodal local-global fusion module is proposed, which consists two levels, i.e., the local attention level and global attention level. Finally, the multimodal fusion signals is fed to fully connected layer and softmax circuit for affective video content analysis.

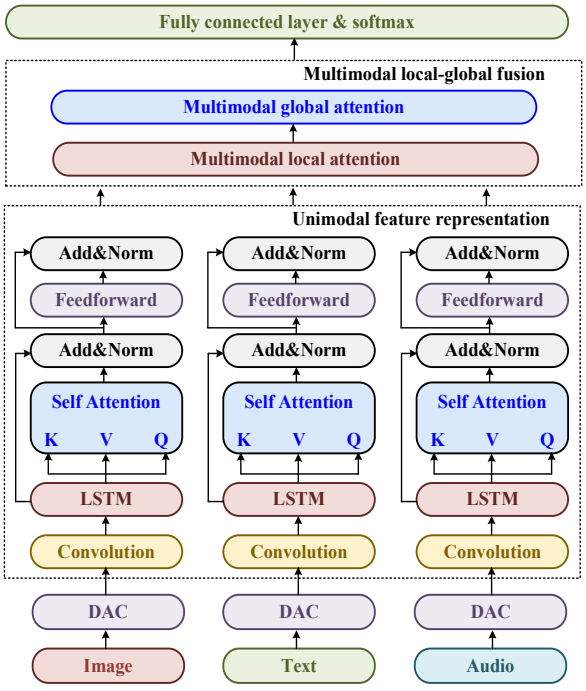

Fig. 1. Overall circuit architecture

B. Circuit Design of Unimodal Feature Representation Module

In this part, we proposed a unimodal feature representation module with cascaded configuration to capture unique characteristics from multimodal signals. The structure of the proposed unimodal feature representation module is illustrated in Fig. 2.

Notably, considering there are existing circuit design implementations for the layer normalization circuit, and ReLU circuit [16], this work mainly focuses on the investigation of the circuit design of the convolution unit, long short-term memory (LSTM) circuit, and self-attention unit. The specific description is provided below.

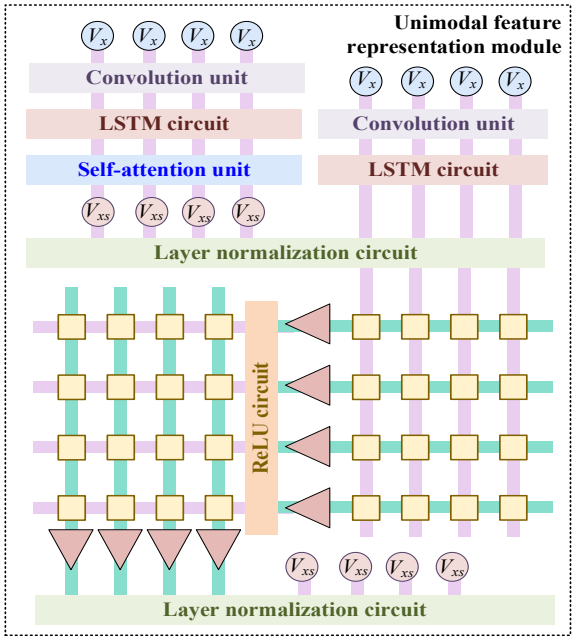

Fig. 2. Circuit design of unimodal feature representation module

1) Circuit design of convolution unit

The convolution unit is designed using the memristor crossbar array with the one-selector–one-Ag/TiO_2/FTO memristor (1S1M) configuration and some peripheral circuits, as shown in Fig. 3. Each column of memristor crossbar array contains M convolution kernels corresponding to an output channel, and the number of columns N is equal to the number of output channels. V_x is the input voltage (x is the index of the modality. x=V, T, A), V_{xc}=Conv(V_x) is the output voltage generated from convolution unit.

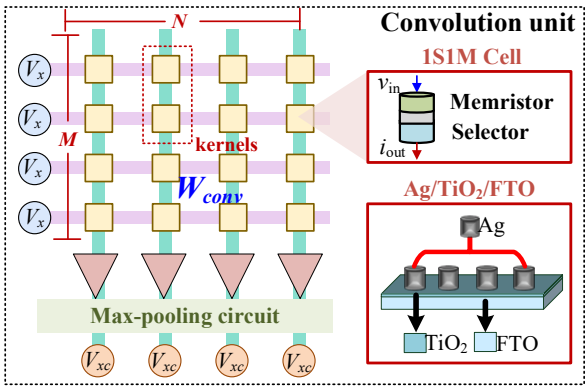

Fig. 3. Circuit design of convolution unit

2) Circuit design of LSTM network

According to [17], the circuit design of LSTM unit is provided in Fig. 4.

From Fig. 4, the LSTM unit is mainly composed by two 1S1M memristor crossbar arrays, one current subtractor [18], and one sigmoid circuit [19]. $V^i_{xc}(t)$ is the input voltage at time step t, $V^h_{xc}(t\text{-}1)$ is the hidden state voltage at time step t-1. V_b is the bias voltage. The weight in each LSTM unit is represented by the difference in conductance of two memristors. The specific input and output of the LSTM unit is provided below:

$$V_{xlu} = \sigma\left(\sum_{t=1}^{T}\left(G_i^+ - G_i^-\right)V_{xc}^i(t) + \sum_{t=1}^{T}\left(G_h^+ - G_h^-\right)V_{xc}^h(t-1) + I_b\right) \quad (1)$$

This article has been accepted for publication in a future proccedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICIT58465.2023.10143111, 2023 IEEE International Conference on Industrial Technology (ICIT)

where σ is sigmoid function, W_i=G^+_i-G^-_i and W_i=G^+_i-G^-_1 are weight matrixes, I_b is the corresponding bias current, V_{xlu} is the output voltage of the LSTM unit.

The LSTM cell mainly consists of four LSTM units to generate input gate voltage, forget gate voltage, previous cell state voltage, and output gate voltage, respectively. Based on this, the circuit design of LSTM network can be obtained, and the output voltage of LSTM network is symbolized by V_{xl}=LSTM(V_{xc}).

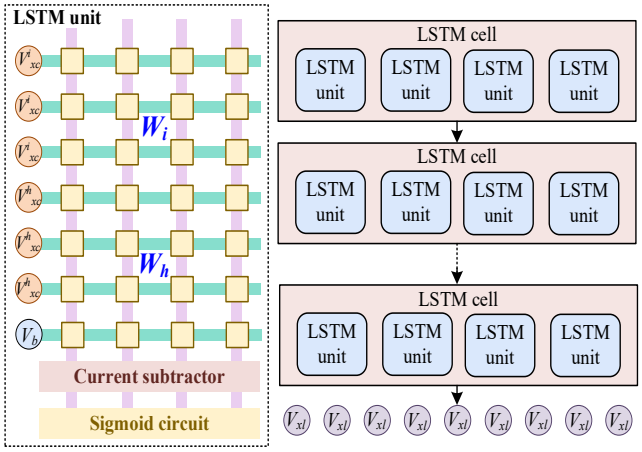

Fig. 4. Circuit design of LSTM unit

3) Circuit design of self-attention unit

The self-attention unit is employed to extract more abundant characteristics, as shown in Fig. 5.

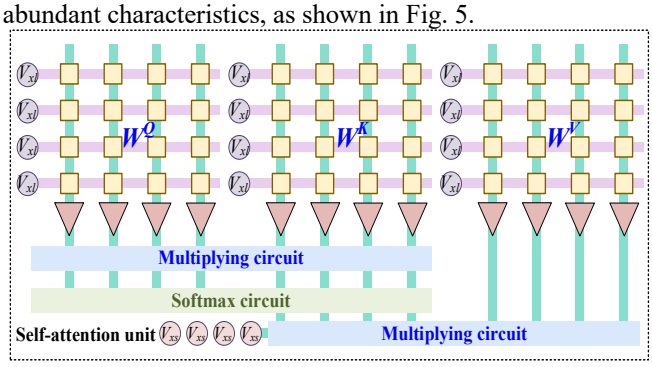

Fig. 5. Circuit design of self-attention unit

From Fig. 5, 1S1M memristor crossbar arrays are used to store the attention weight matrixes W^Q, W^K, and W^V.

The output voltage V_{xs} of self-attention unit is mathematically described by:

$$V_{xs} = soft\max\left(\frac{\left(W^Q V_{xl}\right)^T \cdot \left(W^K V_{xl}\right)^T}{\sqrt{d}}\right) \cdot \left(W^V V_{xl}\right)^T \quad (2)$$

where T means the transpose operation. d is the dimension of the self-attention unit. Following the attention weight matrixes W^Q, W^K, and W^V, the input voltage V_{xl} can be transformed into the attention voltage V_{xs}.

C. Circuit Design of Multimodal Local-Global Fusion Module

In this part, the multimodal local-global fusion module aims at stimulating the process of multimodal information sensing and processing, which consists of two units, i.e., local attention unit and global attention unit, as shown in Fig. 6.

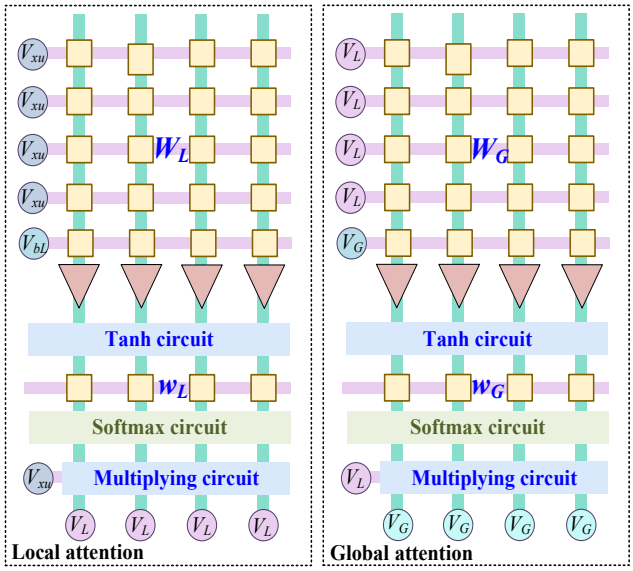

Fig. 6. Circuit design of multimodal local-global fusion module

From Fig. 6, the local attention module is proposed to generate the fusion information which is mainly composed by 1S1M memristor crossbar array, tanh circuit, softmax circuit, and multiplying circuit. The local attention voltage V_L can be obtained by:

$$V_L = \sum soft\max\left(w_L^T \tanh\left(W_L V_{xu} + I_{bL}\right)\right) \cdot V_{xu} \quad (3)$$

where W_L denotes the weigh matrix of local attention. w_L is the parameter vector in local attention unit. It is noted that W_L and w_L are both realized by 1S1M memristor crossbar array. V_{xu} is the input voltage from the unimodal feature representation module. I_{bL} is bias current of the local attention unit.

Considering human tends to focusing the most crucial information in multimodal information processing, the global attention module is designed to capture the key features in the fusion information. The global attention voltage V_G can be mathematically described by:

$$V_G = \sum soft\max\left(w_G^T \tanh\left(W_G V_L + I_{bG}\right)\right) \cdot V_L \quad (4)$$

where W_G denotes the weigh matrix of global attention. w_G is the parameter vector in global attention unit. V_L is the input voltage from the local attention unit. I_{bG} is bias current of the global attention unit.

III. Application in Affective Video Content Analysis

To verify the effectiveness and feasibility, the proposed multimodal attention memristive network is further applied to perform affective video content analysis.

A. Datasets and Evaluation Metrics

Two benchmark affective video datasets (i.e., the LIRIS-ACCEDE dataset and the FilmStim dataset) [20], containing image, text, and audio information, are adopt to demonstrate the evaluation of the proposed network.

Specifically, the LIRIS-ACCEDE dataset is composed by 9800 video excerpts that are extracted and classified for valence and arousal from 160 short movies and feature movies. Valence denotes the happiness ranging from positive to negative, while arousal denotes the emotional intensity ranging from excited to calm. The FilmStim dataset is composed by 70 video excerpts. 50 videos are chosen with a certain emotion label including seven categories, i.e., anger, sadness, disgust, fear, tenderness, neutral, and amusement. For affective video analysis, about 360 participants watched

This article has been accepted for publication in a future proccedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/ICIT58465.2023.10143111, 2023 IEEE International Conference on Industrial Technology (ICIT)

and rated the video, the average rates are used as target value for valence and arousal domain.

Then, common performance metric classification accuracy [21-23] is used to evaluate the overall emotion classification performance.

B. Classification Results

In this paper, the multimodal dates from two benchmark datasets can be converted to the voltage signals (containing image, text, and audio information). Then, these voltage signals are further injected to the proposed multimodal attention memristive network. The output valence voltage V_V and arousal voltage V_A in the two dimensional space repersent human emotion state. Notably, the network parameters obtained by back propagation is directly mapped to the proposed network without read or write to 1S1M crossbar array repeatedly, which can effectively reduce hardware loss. The well-trained multimodal attention memristive network is adopted to perform affective video content analysis.

The proposed network is compared with the state-of-the-art methods on the LIRIS-ACCEDE dataset and the FilmStim dataset, as shown in Table I.

TABLE I. COMPARISION WITH THE STATE-OF-THE-ART RESULTS IN AFFECTIVE VIDEO CONTENT ANALYSIS

Ref.	LIRIS-ACCEDE dataset			FilmStim dataset		
	Valence	Arousal	Time(s)	Valence	Arousal	Time(s)
[4]	43.74	60.88	2321.7	91.88	76.56	227.2
[5]	44.26	60.88	2492.5	90.63	67.19	237.8
[6]	45.82	65.85	2242.8	92.19	70.31	198.6
[7]	44.90	53.11	2379.0	83.87	69.35	210.5
[8]	49.63	64.30	2536.1	94.06	77.19	244.7
Ours	48.22	64.17	102.6	92.41	76.68	11.7

From Table I, the proposed network ranks in the top three on the LIRIS-ACCEDE dataset and FilmStim dataset. For the LIRIS-ACCEDE dataset, the proposed network slightly outperforms state-of-the-art methods in the affective video content analysis task [4-7]. For FilmStim dataset, the proposed network achieves the improvements on accuracy over other competitors. Meanwhile, [8] is slightly superior to the proposed network in terms of accuracy, while inferior to running time. The experimental results demonstrate that the trade-off between classification accuracy and running time can be balanced well in the proposed multimodal memristive network.

To explore the effect among different modalities in classification performance, we use different modalities combinations the LIRIS-ACCEDE dataset and FilmStim dataset, as shown in Table II.

From the classification performance of valence and arousal, the visual modality achieves best accuracy in the two benchmark datasets. The text or audio modality is used together with the visual modality, the classification accuracy slightly outperforming the visual modality. The experimental result demonstrate that the text or audio modality can provide additional information with the audio modality. It is noted that the best classification performance is realized when all modalities are used in the proposed network for affective video content analysis task.

TABLE II. COMPARISION OF DIFFERENT MODALITY FOR AFFECTIVE VIDEO CONTENT ANALYSIS

Modality	LIRIS-ACCEDE dataset		FilmStim dataset	
	Valence	Arousal	Valence	Arousal
T	38.61	55.82	79.32	66.26
A	42.33	58.73	82.54	70.58
V	46.24	61.98	87.63	73.34
T+A	43.45	59.74	84.11	71.96
T+V	47.03	62.12	89.02	74.83
A+V	47.95	63.64	91.32	75.32
A+T+V	48.22	64.17	92.41	76.68

Note: T, A, V represent the textual, audio and visual modalities, respectively.

IV. CONCLUSION

This paper investigates a multimodal attention memristive network for affective video content analysis. Specifically, the proposed network mainly contains unimodal feature representation module and multimodal local-global fusion module. Through the unimodal feature representation module, unique characteristics from different modalities can be adequately extracted. Through the multimodal local-global fusion module select key information both from local and global steams. For verification, the multimodal attention memristive network is applied to perform affective video content analysis. The experimental results demonstrate that the proposed network has good performance in terms of classification accuracy and running time (approximately 10~15 times speed up).

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62001149, Natural Science Foundation of Zhejiang Province under Grant LQ21F010009, and Fundamental Research Funds for the Provincial University of Zhejiang under Grant GK229909299001-06.

REFERENCES

[1] Y. Zhu, Z. Chen and F. Wu, "Affective video content analysis via multimodal deep quality embedding network," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1401-1415, 2022.

[2] Y. Yi, H. Wang and Q. Li, "Affective video content analysis with adaptive fusion recurrent network," IEEE Transactions on Multimedia, vol. 22, no. 9, pp. 2454-2466, 2020.

[3] S. Zhang, C. Yin and Z. Yin, "Multimodal sentiment recognition with multi-task learning," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 7, no. 1, pp. 200-209, 2023.

[4] S. Wang, C. Wang, T. Chen, Y. Wang, Y. Shu and Q. Ji, "Video affective content analysis by exploring domain knowledge," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 1002-1017, 2021.

[5] Q. Gan, S. Wang, L. Hao and Q. Ji, "A multimodal deep regression bayesian network for affective video content analyses," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5123-5132, 2017.

[6] Y. Yi and H. Wang, "Multi-modal learning for affective content analysis in movies," Multimedia Tools and Applications, vol. 78, pp. 13331-13350, 2019.

[7] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo, "Deep sentiment features of context and faces for affective video analysis," Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 72-77, 2017.

[8] Y. Ou, Z. Chen and F. Wu, "Multimodal local-global attention Network for affective video content analysis," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1901-1914, 2021.

[9] X. Ji, C. S. Lai, G. Zhou, Z. Dong, D. Qi and L. L. Lai, "A flexible memristor model with electronic resistive switching memory behavior and its application in spiking neural network," IEEE Transactions on NanoBioscience, vol. 22, no. 1, pp. 52-62, 2023.

[10] X. Ji, D. Qi, Z. Dong, et al. TSSM: Three-State Switchable Memristor Model Based on Ag/TiO x Nanobelt/Ti Configuration[J]. International Journal of Bifurcation and Chaos, 2021, 31(07): 2130020.

[11] Z. Dong, X. Ji, C. S. Lai and D. Qi, "Design and implementation of a flexible neuromorphic computing system for affective communication via memristive circuits," IEEE Communications Magazine, vol. 61, no. 1, pp. 74-80, 2023.

[12] Z. Dong, X. Ji, C. S. Lai, D. Qi, G. Zhou and L. L. Lai, "Memristor-based hierarchical attention network for multimodal affective computing in mental health monitoring," IEEE Consumer Electronics Magazine, Early Access, 2022.

[13] Z. Dong, X. Ji, G. Zhou, M. Gao and D. Qi, "Multimodal neuromorphic sensory-processing system with memristor circuits for smart home applications," IEEE Transactions on Industry Applications, vol. 59, no. 1, pp. 47-58, 2023.

[14] X. Ji, Z. Dong, C. S. Lai and D. Qi, "A brain-inspired in-memory computing system for neuronal communication via memristive circuits," IEEE Communications Magazine, vol. 60, no. 1, pp. 100-106, 2022.

[15] X. Ji, Z. Dong, C. S. Lai, G. Zhou, and D. Qi, " A physics-oriented memristor model with the coexistence of NDR effect and RS memory behavior for bio-inspired computing," Materials Today Advances, vol. 16, pp.100293, 2022.

[16] C. Yang, X. Wang and Z. Zeng, "Full-circuit implementation of transformer network based on memristor," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 4, pp. 1395-1407, 2022.

[17] S. Wen et al., "memristive LSTM network for sentiment analysis," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 3, pp. 1794-1804, 2021.

[18] N. Bansal and R. Pandey, "A Novel Current Subtractor Based on Modified Wilson Current Mirror Using PMOS Transistors," 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), 2016, pp. 444-449.

[19] F. Liu, B. Zhang, G. Chen, G. Gong, H. Lu and W. Li, "A Novel Configurable High-precision and Low-cost Circuit Design of Sigmoid and Tanh Activation Function," 2021 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), Zhuhai, China, pp. 222-223, 2021.

[20] Y. Baveye, E. Dellandréa, C. Chamaret and L. Chen, "LIRIS-ACCEDE: A Video Database for Affective Content Analysis," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 43-55, 2015.

[21] B. Dudzik, H. Hung, M. A. Neerincx and J. Broekens, "Collecting Mementos: A Multimodal Dataset for Context-Sensitive Modeling of Affect and Memory Processing in Responses to Videos," IEEE Transactions on Affective Computing, Early Access, 2022.

[22] Z. Zhang, Z. Dong, H. Lin, et al. An improved bidirectional gated recurrent unit method for accurate state-of-charge estimation[J]. IEEE Access, 2021, 9: 11252-11263.

[23] M. Wang, Y. He, X. Xu, et al. A review of AC and DC electric springs[J]. IEEE Access, 2021, 9: 14398-14408.