Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **2**, May **2025**

Multimodal Fusion Visual Communication Method Based on Genetic Algorithm

Jun-Jie Cao*

Jing'an Branch, Shanghai Open University, Shanghai 200040, P. R. China
junjie6169@163.com

Si-Min Fang

Information Center, Shanghai Jing'an District College, Shanghai 200040, P. R. China
cyberfang601@126.com

Harley Contreras

IS Support, Exiger Canada Inc., Toronto, M5E 1G6, Canada
Harleycontre88@gmail.com

*Corresponding author: Jun-Jie Cao

Received May 15, 2024; revised September 4, 2024; accepted December 16, 2024.

Abstract. *Aiming at the issues of weak correlation between image features and text features and poor visual communication effect in current visual communication methods, this article researches the multimodal fusion visual communication method based on Genetic Algorithm (GA). Firstly, the crossover rate and variation rate of the GA are improved, and a linear function is adopted for adaptive adjustment, which effectively solves the problems of early maturity. Secondly, TextCNN and BERT models are adopted to extract textual features of visual information, and residual networks are used to extract image features of visual information; then, the cross-modal attention mechanism is used to realize inter-modal information interaction, and important information between textual modality and image modality is obtained; meanwhile, the self-attention mechanism is used to fuse the information within the modalities to suppress noise interference. Finally, the fused features are used as the input of Retinex visual communication method, the original Gaussian function is replaced by the improved bilateral filter, and the spatial distance difference scale factor is optimized by the improved GA to explore the hyperparameter space more comprehensively, to improve the efficiency and quality of Retinex algorithm, and to realize the enhancement of visual information. The experimental outcome indicates that the suggested method is better than the other two methods in terms of discrete entropy, sharpness and contrast.*

Keywords: Visual communication; Genetic algorithm; Multimodal fusion; Attention mechanism; Retinex algorithm

1. **Introduction.** In the age of Internet, all kinds of visual media are conveying all kinds of information to people. Visual communication, as an emerging discipline, summarizes and researches the communication mode of visual language [1]. In the field of visual communication, pictures are an important carrier of information transmission, and high-quality pictures can transmit more complete and effective information and improve the transmission effect. People rely more and more on visual information to get and understand the content. The position of visual communication in information dissemination is

becoming more and more prominent, and it is of great significance to improve the communication effect and satisfy the users' needs [2, 3]. However, in the process of generating and transmitting pictures, the quality of pictures can be reduced due to various external unfavorable factors. At this time, it is necessary to enhance the picture to ensure the visual effect of the picture [4]. The traditional visual communication method lacks the support of modern image processing technology, which leads to the blurring of the enhanced images [5]. Therefore, designing an efficient visual communication method to realize visual enhancement is a hot research topic at present.

1.1. **Related work.** Huang et al. [6] introduced the visual communication technique to adjust the brightness of a local area to enhance the image, Venu and Anuradha [7] used the visual communication technique to classify the image into spatial and temporal, intrinsic and extrinsic, and abstract and concrete, and then extracted the image features, and then projected the image features into a clustering matrix for enhancement. Hayat and Imran [8] used the luminance mapping function to complete the enhancement process by transforming the color space model and combining the exposure interpolation method with the multiscale fusion strategy to improve the visual quality. Vijilin and Govindan [9] proposed a visual optimization method based on the improved wavelet thresholding function, but the efficiency of image processing is not high. Bai et al. [10] obtained the initial grayscale image by solving the global mapping function, and then scale decomposition, and combined with the contour wavelet transform to obtain the enhanced image. Bhandari et al. [11] introduced the Gamma transform to process the details of the image, and then fused the processed images to obtain the final enhanced image.

Kuang et al. [12] input the acquired low-frequency information into a convolutional neural network and reconstructed the high-frequency information of the video image on the basis of it, and realized the visual communication of the video image through the enhanced visual effect, but the visual image was blurred. Yan et al. [13] introduced a relative mean generative adversarial network reinforcement learning framework and used the objective function in the actor-critic algorithm as a penalty strategy in the gradient algorithm to improve the image enhancement ability of the model. Matin et al. [14] firstly established a visual communication Retinex model, enhanced the brightness of the image through adaptive normalization function; finally, combined with the Particle Swarm Optimization (PSO) algorithm to highlight the details of the image, to achieve the optimization of the visual communication effect. Pramanik [15] denoised the image based on wavelet decomposition and used GA for optimization enhancement of visually communicated image contrast, but the visual quality was poor.

With the popularization of informationization and digitization in the society, the visual forms we face have also changed, and the most important change is that the dissemination of visual information has changed from unimodal to multimodal. Ji et al. [16] also proposed a memory fusion network for the interaction of image modality and text modality, and mined the single-view features by using the LSTM, but there is a certain loss of the detail information, which affects the visual effect. Lv et al. [17] augmented image RGB features and text features with modal attention to aggregate branches to obtain the aggregated features. Wu et al. [18] used BERT and attention mechanism to mine the association between image modality and text modality. Wang et al. [19] introduced the convolutional attention module to enhance the visual feature extraction, and finally the Retinex algorithm was optimized by PSO for image enhancement.

1.2. **Contribution.** In view of the issues of insufficient multimodal feature fusion and poor visual communication effect in the current visual communication methods, this article suggests a multimodal fusion visual communication method based on GA to obtain high-quality visual information and meet the visual requirements. Firstly, the single-point crossover and mutation operators in the GA are improved to effectively increase the convergence speed of the optimal solution. Secondly, TextCNN and BERT models are used to extract text local features and text context features of the visual information, and residual networks are used to extract image features of the visual information; then, the attention mechanism is introduced into the multi-channel feature extraction and fusion process to obtain the important information between text modality and image modality, and the features are fused to suppress the noise interference. Finally, the fused features are used as the input of Retinex visual communication method, and the improved GA is used to iteratively search for the optimization of the spatial distance difference scale factor, so as to improve the efficiency and quality of the Retinex algorithm and realize the enhancement of visual information. Simulation results show that the proposed method greatly improves the signal-to-noise ratio and average gradient value of graphic visual information, and the visual communication effect is better.

2. Theoretical analysis.

2.1. Multimodal feature fusion. Multimodal feature fusion [20] refers to the fusion of modal features at the modal level, where modal data are first extracted or constructed by simple feature extraction, and then the original input features are converted into higher-level joint features by an intermediate layer of the neural network. In fusion, after extracting the features of different modalities, the features of different modalities can be fused into a single hidden layer to train the network to learn the multimodal joint features. This is implied in Figure 1.

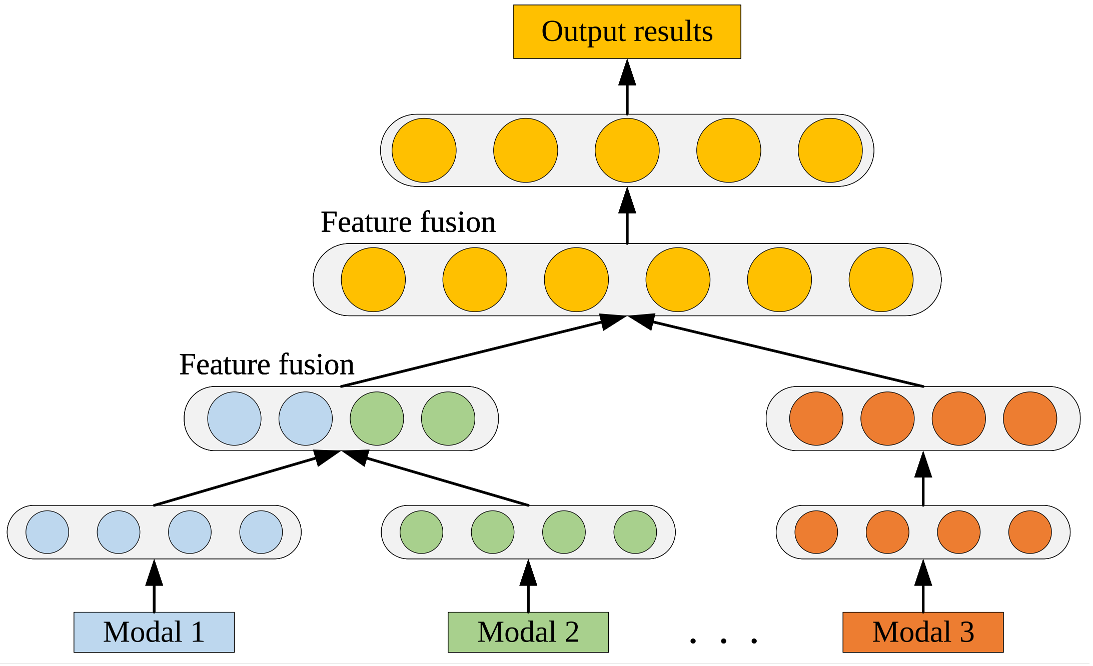

Figure 1. Multimodal feature fusion

The usual methods for feature fusion are Element-wise Addition, Concatenate, and Element-wise Multiplication [21]. The multimodal decomposition higher-order pooling method represents the multimodal data as a low-rank tensor that contains semantic information about each modality and the correlations between modalities. This low-rank tensor is then dimensionalized using a higher-order pooling operation to obtain a compact representation. Higher-order semantic correlations can be learned and model performance can be improved using this approach.

2.2. **Retinex algorithm.** The Retinex algorithm belongs to the group of visual communication technologies [22], which performs visual enhancement based on the simulation of the processes in the human retina and cerebral cortex, which can be represented as bellow.

$$D(x, y) = E(x, y) \times K(x, y) \tag{1}$$

where $D(x, y)$ is the image after brightness correction; $K(x, y)$ is the light component in the image; $E(x, y)$ is the reflection image; and $E(x, y)$ can reflect the image's detail information.

The core idea of the Retinex algorithm is to remove the influence of $K(x, y)$ on the visual effect of the image and retain $E(x, y)$, which reflects the essential properties of the image. The single-size Retinex algorithm analyzes the lighting component $K(x, y)$ by means of a Gaussian function to eliminate the effect of $K(x, y)$ on the real-world image to enhance the image, as described below.

$$E(x, y) = \ln D(x, y) - \ln[\mu(x, y) * D(x, y)] \tag{2}$$

where the convolution operation is described by *; the Gaussian kernel function is described by $\mu(x, y)$ and is expressed as bellow.

$$\mu(x, y) = \xi \times e^{-(x^2+y^2)/\theta^2} \tag{3}$$

where the normalization factor is described by ξ; the Gaussian wrap-around scale is described by θ; and e represents the exponential function.

In the Retinex algorithm, the value of θ is crucial, which is directly related to the visual enhancement effect of the image [23]. Only when the value of θ is appropriate, can the enhanced image maintain a balance between dynamic range compression and contrast, and achieve the ideal image enhancement effect. In order to make the value of θ more reasonable, this article will utilize GA to select the value of θ adaptively.

3. **Optimization of genetic algorithm.** In traditional GA, the genetic operators use constant crossover probabilities and mutation probabilities, which are prone to the problem of "early maturation", leading to the final result falling into local optimization [24]. In addition, the selection, crossover and mutation operations in the GA, respectively, determine the offspring individuals under a certain probability, which may lead to the optimal individuals in the population being "screened out" as the evolutionary process proceeds, so that the optimal values in the contemporary population cannot be inherited to the next generation, thus affecting the convergence of the algorithm to the optimal solution.

Intending to the above issues, this article firstly improves the crossover rate and variation rate, the crossover probability and variation probability can be changed at the right time, and a linear function is adopted for adaptive change adjustment, which effectively solves the issues of early maturity. The formulas of crossover rate and variation rate are implied in Equation (4) and Equation (5).

$$P_c = \begin{cases} c_1 - \frac{(c_1-c_2)(f'-f_{avg})}{f_{max}-f_{avg}} & f' \geq f_{avg} \\ c_1 & \text{other} \end{cases} \tag{4}$$

$$P_m = \begin{cases} m_1 - \frac{(m_1-m_2)(f-f_{avg})}{f_{max}-f_{avg}} & f \geq f_{avg} \\ m_1 & \text{other} \end{cases} \tag{5}$$

where c_1 and c_2 are the maximum and minimum values of crossover rate; m_1 and m_2 are the maximum and minimum values of mutation rate; f_{max} and f_{avg} are the maximum fitness of the individuals in the population and the average fitness of all individuals, respectively; f is the fitness of the individual that is about to be mutated in the population; f' is the larger fitness of the two individuals during the crossover operation.

Intuitively, the above formula adjusts the mutation rate and crossover rate linearly, instead of fixing them. When the fitness of an individual calculated from the fitness function is lower than the average fitness, it means that the solution represented by that individual is less effective, and then, according to the idea of genetic algorithms, it is evolved to be larger, i.e., the crossover rate and the mutation rate are larger. If the fitness of individuals in the population is high, then linear adjustment is performed.

Then for the issue of slow convergence, this paper adopts the following GA improvement strategy.

(1) Before GA in the l-th generation of the parent population, calculate the fitness function $F_{xl}^1, F_{xl}^2, \ldots, F_{xl}^q$ corresponding to q individuals $x_l^1, x_l^2, \ldots, x_l^q$ (q is the population size), and select the top m best individuals A_{ox} in the parent population.

$$A_{ox} = \left(x_l^{i_1}, x_l^{i_2}, ..., x_l^{i_m}\right) \tag{6}$$

$$(i_1, i_2, ..., i_m) = \max\left\{F_{xl}^1, F_{xl}^2, ..., F_{xl}^m\right\} \tag{7}$$

where $i_1, i_2, ..., i_m$ is the m best individual index markers before evolution.

Secondly, according to the evolutionary process of GA, the parent population $x_l^1, x_l^2, ..., x_l^q$ evolved to the offspring population $y_l^1, y_l^2, ..., y_l^q$.

Then, the fitness function value $F_{yl}^1, F_{yl}^2, ..., F_{yl}^q$ is calculated for q offspring individuals $y_l^1, y_l^2, ..., y_l^q$. In the lth generation of the offspring population, the n worst individuals A_{oy} of the offspring population are selected.

$$A_{oy} = \left(y_l^{j_1}, y_l^{j_2}, ..., y_l^{j_m}\right) \tag{8}$$

$$(j_1, j_2, ..., j_m) = \min\left\{F_{yl}^1, F_{yl}^2, ..., F_{yl}^m\right\} \tag{9}$$

where $j_1, j_2, ..., j_m$ is the m worst individual index markers in the offspring population.

Finally, in the offspring population individual $y_l^1, y_l^2, ..., y_l^q$, A_{ox} is replaced by A_{oy} to obtain the parent population individual $x_{l+1}^1, x_{l+1}^2, ..., x_{l+1}^q$ in the next generation.

4. Research on multimodal fusion visual communication method based on GA.

4.1. Feature extraction of visual information.

Focusing on the issues of poor image quality and missing text information in current visual communication methods, this paper designs a multi-modal fusion visual communication method based on GA. Firstly, TextCNN and BERT models are used to extract text local features and text context features respectively, and residual networks are used to extract image features; secondly, a cross-modal attention mechanism is introduced to realize the information interaction between graphic and text modalities. Through this cross-modal attention mechanism, the important information between text modality and image modality can be obtained from each other, and at the same time, the information fusion is carried out within the modality by using the self-attention mechanism to inhibit the noise interference and reduce the information redundancy. Finally, the GA-optimized Retinex algorithm is used to enhance the graphic and textual visual information to achieve the optimization of visual communication effect. The overall model is shown in Figure 2.

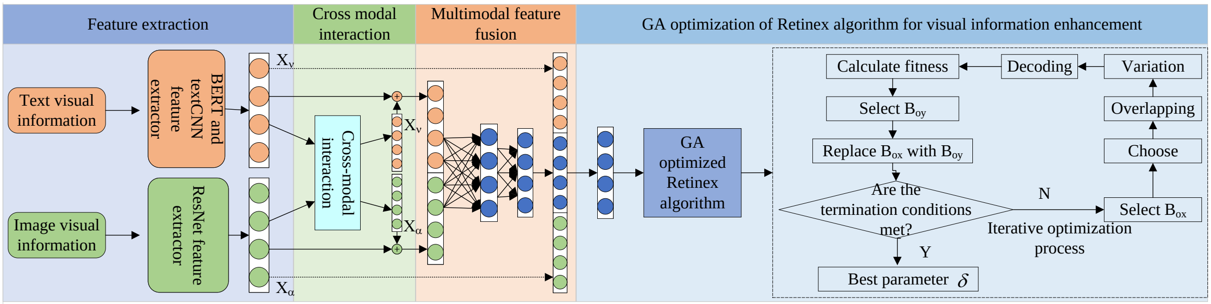

Figure 2. GA optimized Retinex algorithm for visual information enhancement models

Existing visual information visible to our naked eyes includes images and text, in this paper, this article firstly performs feature extraction for image modality and text modality in visual information.

Given a graphic pair (I, S), for text modal, and the text consists of m words, the text description is denoted as $S = \{s_1, s_2, ..., s_m\}$. The text context features are extracted using BERT [25] to obtain the word embedding representation of the text, as implied below.

$$E = \{e_1, e_2, ..., e_m\} = BERT(S) \tag{10}$$

Secondly, the obtained word embedding representation is inputted into the coding layer for training, and finally the contextual feature representation of the text is obtained, as shown in Equation (11).

$$S^t = BERT(\{e_1, e_2, ..., e_n\}) \tag{11}$$

To fully capture the visual text features, TextCNN is used to extract the local features of the text. Firstly, the word embedding technique is used to represent each word as a d-dimensional word vector $s_i^v \in R^d$, in which the text after word embedding is $S_i = \{s_1^v, s_2^v, ..., s_m^v\}$. Then, TextCNN [26] is adopted to extract the local features of the text, as implied bellow.

$$S^q = TextCNN(\{s_1^v, s_2^v, ..., s_m^v\}) \tag{12}$$

For the image modality, the pre-trained ResNet is used for image feature extraction, and the obtained image features are represented as indicated in Equation (13).

$$I^t = ResNet(I) \tag{13}$$

where $I^t \in R^{width*height*fm}$, $width$ represents the width of the image, $height$ represents the height of the image, and fm represents the number of feature maps. The values of $width$, $height$ and fm are different for different output layers of ResNet.

4.2. **Feature fusion for multimodal visual information.** Firstly, the cross-modal attention mechanism is used to make text and image features guide each other and capture the inter-modal correlation, so as to realize the information interaction between different modalities. When there are two modal features α and β, denoted as $X_\alpha \in R^{S_\alpha \times d_\alpha}$ and $X_\beta \in R^{S_\beta \times d_\beta}$; the query vector is denoted as $Q_\alpha = X_\alpha V_{Q_\alpha}$, the key vector is denoted as $K_\beta = X_\beta V_{K_\beta}$, the value vector is denoted as $W_\beta = X_\beta V_{W_\beta}$, and V is the trainable parameter matrix. The α-guided β-modal features are denoted as bellow.

$$CA_{\alpha\to\beta}\left(X_\alpha, X_\beta\right) = soft\max\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right) \tag{14}$$

$$W_\beta = soft\max\left(\frac{X_\alpha V_{Q_\alpha} V_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}\right) X_\beta V_{W_\beta} \tag{15}$$

Input image features as *Query*, text context features as *Key* and *Value* into cross-modal attention network to get image-guided text context features $Z_{st}^{I^t}$. Input image features as *Query*, text local features as *Key* and *Value* into cross-modal attention network to get image-guided text local features $Z_{sq}^{I^t}$. Similarly, get image features $Z_{It}^{S^t}$ guided by text context features and image features Z_{It}^{Sq} guided by text local features.

Then the four groups of cross-modal interaction features are spliced two by two to obtain the image-text association feature Z_S, text-image association feature Z_I, respectively, as implied below.

$$Z_S = Concat\left[Z_S^{I^t}, Z_S^{I^q}\right] \tag{16}$$

$$Z_I = Concat\left[Z_I^{S^t}, Z_I^{S^q}\right] \tag{17}$$

Feature fusion is performed by self-attention mechanism to obtain the internal correlation of image and text association features. Z_S and Z_I are spliced to get Z', and input into Equation (18) to get Equation (19).

$$Self - Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{18}$$

$$\hat{Z} = self - Attention(Z') \tag{19}$$

where $\hat{Z}$ is the final fused total characteristic of the modes.

4.3. **Visual information enhancement based on GA optimized Retinex algorithm.** In this article, $(I, S, \hat{Z})$ is used as the input for the visual information enhancement of Retinex algorithm, and the Gaussian function is used to process the image, which will cause the enhanced image to have the problem of unclear edges and lack of local information. Therefore, a modified bilateral filter [27] is used to replace the original Gaussian function, and the visual information enhancement is realized by reconstructing the obtained image. The bilateral filter contains two functions, which can be described by the following equation.

$$f'(i,j,k) = \frac{\sum_{(I,S,\hat{Z})\in\Omega_{p,i,j,k}} g_d(I,S,\hat{Z}) g_r(I,S,\hat{Z}) I(I,S,\hat{Z})}{\sum_{(I,S,\hat{Z})\in\Omega_{p,i,j,k}} g_d(I,S,\hat{Z}) g_r(I,S,\hat{Z})} \tag{20}$$

$$w_d(I,S,\hat{Z}) = \exp[-2^{-1}\delta_d^{-2}[(i-I)^2 + (j-S)^2 + (k-\hat{Z})^2]] \tag{21}$$

$$w_r(I,S,\hat{Z}) = \exp\{-2^{-1}\delta_r^{-2}[f(i,j,k) - f(I,S,\hat{Z})]^2\} \tag{22}$$

where the input graphic of the bilateral filter is $J(I,S,\hat{Z})$, and its output is $f'(i,j,k)$. The spatial domain kernel function is denoted by $g_d(I,S,\hat{Z})$, and the spatial distance scale factor is δ. For the graphic $J(I,S,\hat{Z})$, the center of mass (i,j,k) is the set of graphic data contained in the region, denoted by $\Omega_{p,i,j,k}$. The filtering step is p, whose value is directly proportional to the filtering interval, and when it takes a larger value, it means that the difficulty of the operation is also higher.

The value of parameter δ has a direct impact on the processing effect of the bilateral filter, so this paper utilizes the improved GA to find the optimal value of δ. The steps in detail are as bellow.

(1) Chromosome coding and population initialization. Chromosomes are coded by real number coding, the length of chromosome coding is the number of hyperparameters, and the initial value of chromosome is set to be equal to the random point in the search space of hyperparameters. Each set of hyperparameters $(x_1, x_2, ..., x_m)$ is configured as follows:

$$minBd_i \leq x_i \leq maxBd_i; x_i \in R, i = 1, 2, ..., m \tag{23}$$

where x_i is the i-th hyperparameter; $minBd_i$ and $maxBd_i$ are the upper and lower bounds of the i-th hyperparameter x_i.

(2) Adaptation function configuration. When constraining and optimizing the spatial distance difference scale factor of the bilateral filter, the parameter fitness function is constructed in the way of the minimum error time integration objective, so that the application of optimized parameters can have good dynamic process performance. The way for calculating the spatial distance difference scale factor of the optimized system is as follows:

$$\delta = o_1 e(x) - \lambda(x) \tag{24}$$

where δ is the parameter optimized by the GA; o_1 is the output of the bilateral filter; and λ is the control energy.

To avoid the phenomenon of parameter overshooting, the optimized parameters are trained and the optimized parameter δ fitness function $f = 1/\delta$ is constructed.

(3) Crossover and mutation operations in GA are the key to maintain good status and diversity of the population. The crossover operation formula for chromosome i, X_i, and chromosome j, Y_j, at position k after the improvement of GA is as bellow.

If the k-th gene X_i, k of the i-th chromosome is selected for the mutation operation, the mutation operator is as bellow.

$$X_{i,k} = \begin{cases} X_{i,k} + (X_{max} - X_{i,k})(1 - t/T_{max}), & r \geq 0.5 \\ X_{i,k} + (X_{min} - X_{i,k})(1 - t/T_{max}), & r < 0.5 \end{cases} \tag{25}$$

where X_{max} and X_{min} are the upper and lower limits of gene X, respectively; r is a random number in $[0, 1]$, which determines the direction of chromosome variation; t is the current number of iterations; and T_{max} is the maximum number of evolutions.

(4) Parametric chaotic optimization. Collect the sampling data of the filter, and according to the sampling data and sampling frequency, carry out the parameter δ chaotic optimization, and the function expression of this process is as bellow.

$$U(t) = k\left[f + \frac{1}{T} + \int e(t)dt\right] \tag{26}$$

where U is the parametric chaotic optimization; k is the sampled data, $e(t) = (\delta + \lambda(t))/o_1$.

To ensure that the sampled data can play the expected effect and optimal role in parameter chaotic optimization, it is necessary to optimize the sampled data with the following equation.

$$k = \eta(1 - k_1) \tag{27}$$

where η is the optimization of sampling data; k_1 is the sampling data at different points in time. Substituting the sampling data calculated in Equation (27) into Equation (26), the optimal values of the parameters can be output.

5. **Performance testing and analysis.**

5.1. **Quantitative analysis of visual communication effects.**

To estimate the effectiveness of the proposed multimodal fusion visual communication method based on genetic algorithm, 13688 visual information containing graphic data are selected from the multiview multi-source dataset [28], and compared with the WTGA method in the literature [15] and the MSTA method in the literature [19]. The dataset is divided into the training set, testing set and validation set according to 5:3:2. The experimental environment is: Windows 10 system, Intel(R) Core(TM) i3-2120 CPU with 2.30GHz, 4.00GByte of RAM, and Matlab software version 2015a for experimental image processing.

In this article, the contrast CR, discrete entropy HP, and clarity QV [29] are used comprehensively to compare the visual communication effects of the three methods in the experiments. The experimental results are shown in Table 1, and the proposed method is better than WTGA and MSTA in terms of visual discrete entropy, clarity, and contrast. Although the WTGA method utilizes wavelet decomposition to denoise the image, it does not consider the multimodal features of vision, which leads to its worst visual communication effect. And although MSTA considers the multimodal visual information of the graphic in the feature extraction, it does not optimize the traditional genetic algorithm and does not search for the optimization of the relevant parameters of the visual communication technology, which results in its visual communication effect being worse than that of ours method. The CR, HP, and QV of ours method are 7.91, 14.38, and 3.49, respectively, which are enhanced by 2.21, 10.08, and 2.74 compared to WTGA, and 0.56, 3.41, and 1.35 compared to MSTA. From the experimental values, WTGA, MSTA and ours algorithms all enhance the original graphical visual information to different degrees, but the ours method is comprehensively better than the other two methods in terms of discrete entropy, clarity, and contrast indicators.

Table 1. Time-consuming test results for visual information enhancement

Method	**CR**	**HP**	**QV**
WTGA	5.7	4.3	0.75
MSTA	7.35	10.97	2.14
ours	7.91	14.38	3.49

Based on the above experimental results, the time consumption of the above three methods in visual information enhancement is tested, and the test results are implied in Table 2. The analysis of Table 2 shows that the more edge information there is in the graphic, the longer the optimization time is needed to convey the visual information. Among them, ours method can control the optimization time within 50 seconds, which is lower than that of WTGA and MSTA, when the number of samples is increasing and the edge information is getting more and more, which further proves that ours method has a higher processing efficiency when enhancing the communication effect of graphic visual information.

Table 2. Time-consuming test results of for visual information enhancement

The number of samples	WTGA	MSTA	ours
10	25	13	7
20	31	20	15
30	46	32	24
40	64	47	36
50	88	62	45

5.2. Comparison of signal-to-noise ratio and mean gradient for visual communication methods. Taking the signal-to-noise ratio as an index, we use ours method, WTGA method and MSTA method to carry out the visual communication test and compare the signal-to-noise ratios of video images of different methods, and the test results are as follows. Analyzing Figure 3, it can be seen that the SNR of visual information of ours method is higher than that of WTGA and MSTA, which indicates that the larger the SNR is, the better the quality of the image is. WTGA utilizes wavelet decomposition for denoising, but the visual information is redundant, and it is easy to lose a large amount of detail information; MSTA has better denoising effect than WTGA, but the overall clarity of the visual information of the image and text is lower; ours method utilizes self-attention mechanism for information fusion within modalities, and suppresses noise interference, and the overall denoising effect is more natural.

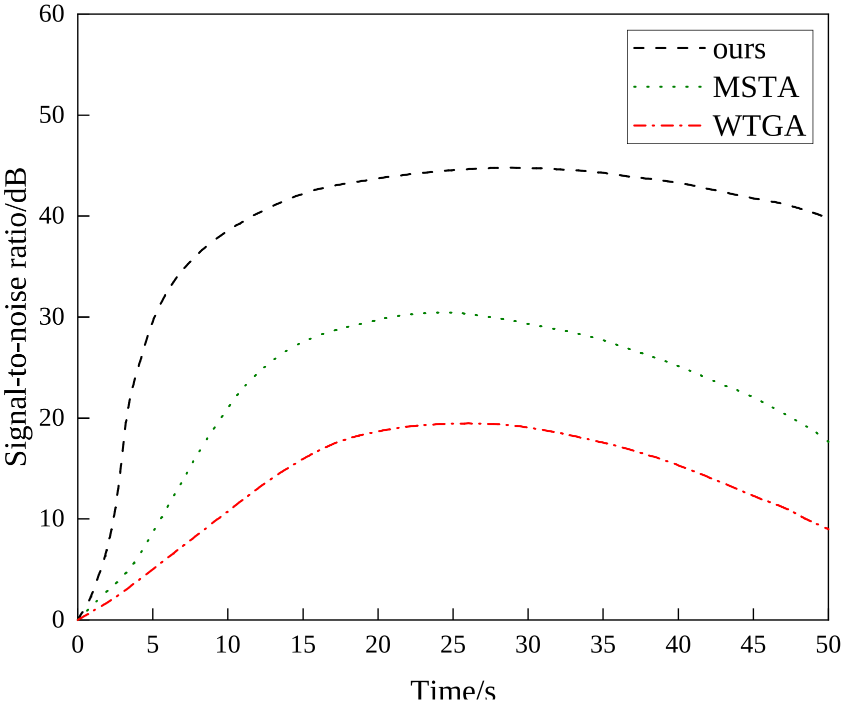

Figure 3. Signal-to-noise ratio test results of the three methods

A comparison of the average gradient of different methods over time is shown in Figure 4. As can be seen from the comparison test results in Figure 4, the ours method not only has much higher index values than the comparison methods, but also has a shorter optimization time than the WTGA and MSTA methods, and it only takes less than 50 seconds to complete the gradient optimization of the parameters. The experimental conclusions obtained show that the ours method uses BERT and TextCNN to extract the graphic and textual features of visual information, and uses the attention mechanism for feature fusion, which reduces the redundancy of information and the difficulty of optimization. In addition, the optimization of the important parameters of Retinex algorithm by using the improved GA significantly increases the average gradient value of the graphic and textual visual, which leads to better visual communication effect and faster processing rate.

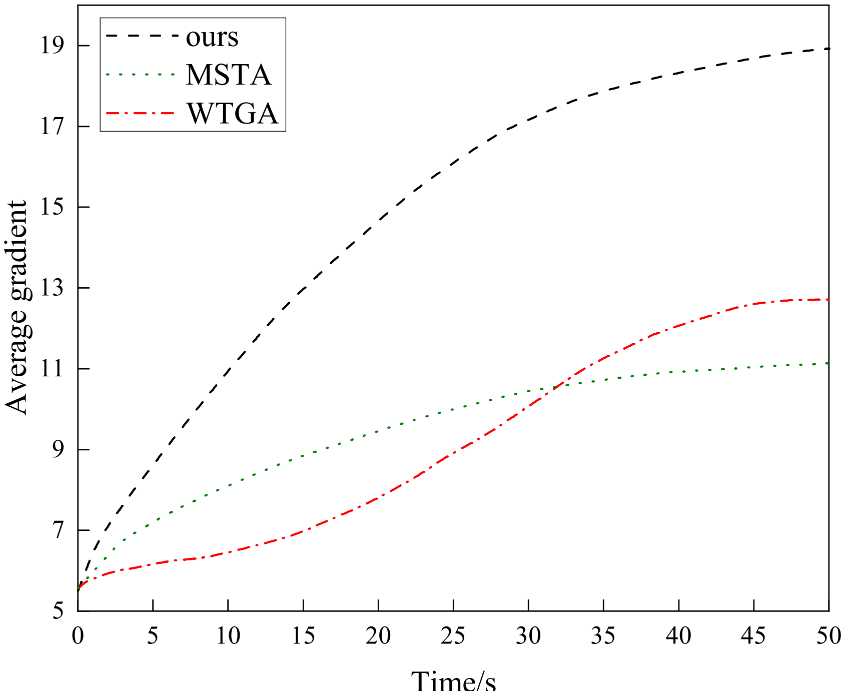

Figure 4. Comparison of the average gradient of the three methods

6. **Conclusion.** In view of the issues of insufficient multimodal feature fusion and poor visual communication effect in current visual communication methods, this article suggests a multimodal fusion visual communication method based on GA. Firstly, the GA is optimized, and a linear function is used for adaptive adjustment, which effectively improves the convergence speed of the optimal solution. Secondly, TextCNN and BERT models are used to extract text local features and text context features of visual information, and residual networks are used to extract image features of visual information; then, cross-modal attention mechanism is used to realize the inter-modal information interaction, and to obtain the important information between text modality and image modality. At the same time, the self-attention mechanism is utilized to fuse the information within modalities to reduce the information redundancy. Finally, the fused features are used as inputs to the Retinex visual communication method, and the improved GA is used to iteratively search for the optimization of the spatial distance difference scale factor, so as to improve the efficiency and quality of the Retinex algorithm, and to realize the enhancement of visual information. The experimental outcome indicates that the designed method improves the discrete entropy, clarity and contrast of visual information, and the processing time is shorter, so it has good practicability and real-time performance.

REFERENCES

[1] D. Osorio, and M. Vorobyev, "A review of the evolution of animal colour vision and visual communication signals," *Vision Research*, vol. 48, no. 20, pp. 2042–2051, 2008.

[2] D. Schill, “The visual image and the political image: A review of visual communication research in the field of political communication,” *Review of Communication*, vol. 12, no. 2, pp. 118–142, 2012.

[3] R. Wang, “Computer-aided interaction of visual communication technology and art in new media scenes,” *Computer-Aided Design and Applications*, vol. 19, no. 3, pp. 75–84, 2021.

[4] J. S. Wolffsohn, D. Mukhopadhyay, and M. Rubinstein, “Image enhancement of real-time television to benefit the visually impaired,” *American Journal of Ophthalmology*, vol. 144, no. 3, pp. 436–440, 2007.

[5] W. Wang, X. Wu, X. Yuan, and Z. Gao, “An experiment-based review of low-light image enhancement methods,” *IEEE Access*, vol. 8, pp. 87884–87917, 2020.

[6] K.-Q. Huang, Q. Wang, and Z.-Y. Wu, “Natural color image enhancement and evaluation algorithm based on human visual system,” *Computer Vision and Image Understanding*, vol. 103, no. 1, pp. 52–63, 2006.

[7] N. Venu, and B. Anuradha, “PSNR Based Fuzzy Clustering Algorithms for MRI Medical Image Segmentation,” *International Journal of Image Processing and Visual Communication*, vol. 2, no. 2, pp. 1–7, 2013.

[8] N. Hayat, and M. Imran, “Detailed and enhanced multi-exposure image fusion using recursive filter,” *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 25067–25088, 2020.

[9] B. Vijilin, and V. Govindan, “Optimal Threshold Selection for Wavelet Transform based on Visual Quality,” *International Journal of Computer Applications*, vol. 975, p. 8887, 2013.

[10] L. Bai, W. Zhang, X. Pan, and C. Zhao, “Underwater image enhancement based on global and local equalization of histogram and dual-image multi-scale fusion,” *IEEE Access*, vol. 8, pp. 128973–128990, 2020.

[11] A. K. Bhandari, A. Kumar, G. K. Singh, and V. Soni, “Dark satellite image enhancement using knee transfer function and gamma correction based on DWT–SVD,” *Multidimensional Systems and Signal Processing*, vol. 27, pp. 453–476, 2016.

[12] X. Kuang, X. Sui, Y. Liu, Q. Chen, and G. Gu, “Single infrared image enhancement using a deep convolutional neural network,” *Neurocomputing*, vol. 332, pp. 119–128, 2019.

[13] L. Yan, J. Fu, C. Wang, Z. Ye, H. Chen, and H. Ling, “Enhanced network optimized generative adversarial network for image enhancement,” *Multimedia Tools and Applications*, vol. 80, pp. 14363–14381, 2021.

[14] F. Matin, Y. Jeong, and H. Park, “Retinex-based image enhancement with particle swarm optimization and multi-objective function,” *IEICE Transactions on Information and Systems*, vol. 103, no. 12, pp. 2721–2724, 2020.

[15] S. Pramanik, “An adaptive image steganography approach depending on integer wavelet transform and genetic algorithm,” *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 34287–34319, 2023.

[16] Z. Ji, Z. Lin, H. Wang, and Y. He, “Multi-modal memory enhancement attention network for image-text matching,” *IEEE Access*, vol. 8, pp. 38438–38447, 2020.

[17] C. Lv, B. Wan, X. Zhou, Y. Sun, J. Hu, J. Zhang, and C. Yan, “CAE-Net: Cross-Modal Attention Enhancement Network for RGB-T Salient Object Detection,” *Electronics*, vol. 12, no. 4, p. 953, 2023.

[18] J. Wu, T. Zhu, X. Zheng, and C. Wang, “Multi-modal sentiment analysis based on interactive attention mechanism,” *Applied Sciences*, vol. 12, no. 16, p. 8174, 2022.

[19] X. Wang, S. Hu, and J. Li, “Improved Retinex algorithm for low illumination image enhancement in the chemical plant area,” *Scientific Reports*, vol. 13, no. 1, p. 21932, 2023.

[20] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, “A Spectral Convolutional Neural Network Model Based on Adaptive Fick’s Law for Hyperspectral Image Classification,” *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.

[21] T.-Y. Wu, A. Shao, and J.-S. Pan, “CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm,” *Mathematics*, vol. 11, no. 10, p. 2339, 2023.

[22] T.-Y. Wu, H. Li, and S.-C. Chu, “CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps,” *Mathematics*, vol. 11, no. 9, p. 1977, 2023.

[23] R. R. Hussein, Y. I. Hamodi, and A. S. Rooa, “Retinex theory for color image enhancement: A systematic review,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, p. 5560, 2019.

[24] P. K. Yadav, and N. Prajapati, “An overview of genetic algorithm and modeling,” *International Journal of Scientific and Research Publications*, vol. 2, no. 9, pp. 1–4, 2012.

[25] P. Ganesh, Y. Chen, X. Lou, M. A. Khan, Y. Yang, H. Sajjad, P. Nakov, D. Chen, and M. Winslett, "Compressing large-scale transformer-based models: A case study on bert," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1061–1080, 2021.

[26] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel TextCNN model," *Neurocomputing*, vol. 363, pp. 366–374, 2019.

[27] R. G. Gavaskar, and K. N. Chaudhury, "Fast adaptive bilateral filtering," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 779–790, 2018.

[28] J. Tao, D. Zhou, F. Liu, and B. Zhu, "Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models," *Pattern Recognition*, vol. 87, pp. 296–316, 2019.

[29] T. Celik, "Spatial entropy-based global and local image contrast enhancement," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5298–5308, 2014.