Hindawi
Advances in Multimedia
Volume 2022, Article ID 8437979, 10 pages
https://doi.org/10.1155/2022/8437979

Research Article

Graphic Perception System for Visually Impaired Groups

Jingzi Wen

School of Visual Arts, Hunan Mass Media Vocational and Technical College, Changsha 410100, China

Correspondence should be addressed to Jingzi Wen; fiona_wjz@163.com

Received 31 March 2022; Revised 19 April 2022; Accepted 12 May 2022; Published 3 June 2022

Academic Editor: Qiangyi Li

In the age of internet, the demand of visually impaired groups to perceive graphic images through tactile sense is becoming stronger and stronger. Image object recognition is a basic task in the field of computer vision. In recent years, deep neural networks have promoted the development of image object recognition. However, existing methods generally have problems of image details' loss and edge refinement, which cannot improve the accuracy rate of object recognition for visually impaired groups. In order to solve this problem, this study proposes a graphic perception system, which improves the attention mechanism. This system mainly consists of three modules: mixing attention module (MAM), enhanced receptive field module (ERFM), and multilevel fusion module (MLAM). MAM can generate better semantic features, which can be used to guide feature fusion in the decoding process, so that the aggregated features can better locate significant objects. ERFM can enrich the context information of low-level features and input the enhanced features into MLAM. MLAM uses the semantic information generated by MAM to guide the fusion of the current decoded features and the low-level features' output by ERFM, and gradually recover boundary details in a cascading manner. Finally, the proposed algorithm is compared with other algorithms on PASCAL VOC and MS-COCO data. Experimental results show that the proposed method can effectively improve the accuracy of graphic object recognition.

1. Introduction

At present, blind people mostly use guide poles or other guide devices to assist their daily life. Although these devices can bring great convenience for the blind to travel, they cannot let the blind perceive the appearance of the objects in front of them. Blind people only know that there is an object in front of them, but they cannot know the exact shape of the object. Information is the basis for people to acquire knowledge and communicate with each other. However, for people with visual disabilities, because of visual impairment, other perceptual abilities, such as hearing, touch, smell, and taste perception, have become the main channels for them to obtain information and explore their surrounding environment.

In recent years, more and more researchers have used modern electronic technology to create assistive devices that can help visually impaired groups in their daily life and learning, for example, braille dot display and screen-reading software. [1]. Among them, the tactile sense is the most important and necessary way for people with visual disabilities to learn and recognize images and graphics [2].

In the past, tactile images especially designed and produced for the blind were mostly made using traditional techniques such as thermoplastic vacuum forming, thermal-sensitive printing, and embossing [3]. We make figures and images with protrusions on the surface that can be touched and perceived by fingers. With the development of computers and related software and hardware technologies, electronic braille spot display was first developed, such as Optacon made by Bliss et al. in 1969 [4]. In recent years, researchers have made progress in the development of excitation-related technologies, such as shape-memory alloys, electromagnetic microcoils, air injection, acoustic radiation pressure, pressure valves, and even ionic conductive polymer gel film (ICPF). Subsequently, electronic haptic image display was successively developed, such as the novel BrailleDis 9000 pin-matrix [5], HyperBraille1 [6], Dot View2 [7], and sheet-type Braille displays [8].

However, most of the previous studies are focused on technical implementation, because the images they display are still expressed in the way of traditional visual cognition. For visually impaired groups, their understanding and recognition of images rely on finger touch, which is fundamentally different from the way that sighted people obtain information through external stimuli. These images based on traditional visual cognition are not efficient enough to provide them with effective and accurate information. Experience, skills, and other factors will affect its perception effect. Therefore, how to design and generate image content with clear semantics and easy for blind people to recognize by touch is a great challenge.

Blind people rely on hearing and touch instead of sight to obtain information. In the face of verbal information, blind people can read for them by the real human voice and a computer-simulated human voice through audiotape or screen-reading software [9]. You can also use braille to sense braille through your fingers and understand information [10]. However, no matter in academic research or commercial market, the problem of how to obtain graphic information by the blind has not been well solved.

In view of the problem that the existing visual-tactile display images cannot effectively provide accurate information for the visually disabled, this study makes a deep study on these problems. The algorithm is compared with other algorithms to verify the significance of the system designed in this study for the visual impairment group in the image recognition and provides a new way for the blind to perceive the world, which has high practical value.

The innovations and contributions of this study are listed as follows:

(1) Mixing attention module (MAM). MAM uses the attention mechanism to enhance the saliency of features from the fifth residual layer, so as to get more attention to the semantic features of significant objects. At the same time, in order to solve the problem that the location information of significant objects is constantly diluted in the decoding process, it is used as the semantic guidance in the whole decoding process, and the feature aggregation in the decoding process is continuously guided to generate a more specific significance map.

(2) Enhanced receptive field module (ERFM), which can process the features from the lower layer. The edge details of low-level features are quite rich but limited by the receptive field, and more global information cannot be obtained. Therefore, ERFM is considered to retain the original edge details while obtaining a larger receptive field and enhancing semantic information.

(3) Multilevel aggregation module (MLAM), which efficiently aggregates features generated from the above two modules, continuously extracts significant parts of features in a cascading manner, and refines the edge details of significant objects. The final saliency map is generated.

This study consists of five main parts: the first part is the introduction, the second part is the factors affecting graphic perception, the third part is the graphic perception algorithm based on improved attention mechanism, the fourth part is the experiment and analysis, and the fifth part is the conclusion; besides, there are abstracts and references.

2. Factors Affecting Graphic Perception

Research studies show that there are many factors that affect the perceptual effect of blind people in the process of learning by touching images. It mainly includes the following three points: (1) user's existing experience, memory, and knowledge; (2) user's touch skills; and (3) display of the readability of images and graphics.

2.1. Existing Experience, Memory, and Knowledge of Objects in the User's Mind. People with visual disabilities are divided into congenital disabilities and acquired disabilities [11]. People with acquired visual impairments used to have a vision and have a better cognitive understanding of features than people with congenital visual impairments. For example, people with acquired visual impairment know the features of the sky, trees, lakes, birds, and when they touch the features of images of these objects again, and they can quickly understand, recognize, and rerecognize them. But the innate visual handicap people's cognition of things is through the description of others and memory. In these descriptions, there may be many abstract visual concepts. The sky is blue and big. Poplars in spring are straight up and green. The water is blue. Birds will fly in the sky and so on. Abstract words are difficult for people with congenital visual disabilities to understand and remember and recognize, and when they touch images, there will be more confusion.

2.2. User's Touch Skills. In the process of perceiving tactile images, the perception of speed of images is mostly related to the perceiver's perception mode. For example, studies have shown that top-to-bottom scanning is more efficient than left-to-right when touching images and graphics [12]. Students with better knowledge of graphics will consciously look for its prominent features, such as acute angles, lines, protrusions, and depressions. However, students with poor image recognition skills often lack systematic methods and just cross walk along contours [13].

2.3. Display Image Availability. The two points mentioned in Sections 2.1 and 2.2 are determined by the user's personal situation and education level. By improving the usability design of tactile images, the effect of tactile images perceived by blind people can be improved. Most of the tactile images used by the blind are simply transformed from visual images, without considering the perceptual characteristics of the sense of touch. As shown in Figure 1, there are many complex lines and regions crossing in the image, and the semantic emphasis of the information is not prominent, which leads to the time-consuming and laborious recognition of such visual-tactile images by the visually impaired and often confused [14].

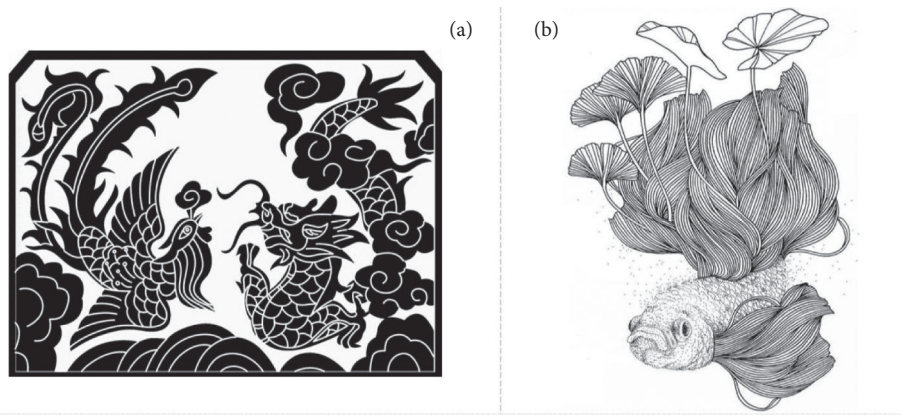

Figure 1: Learning materials of blind students.

In this study, this kind of image is called a "visual-tactile image," referred to as V-image.

At present, in the process of designing electronic tactile displays for visually disabled people, most designers pay attention to the appearance, use environment, and comfort of products, but lack of comprehensive research on the unique characteristics of tactile images. In fact, it is necessary to understand the characteristics of touch, how it is similar and different from vision, and how to design tactile images according to these characteristics before forming the design principles of tactile displays.

3. Graphic Perception Algorithm Based on Improved Attention Mechanism

As shown in Figure 2, a coding-decoding structure is established in this study. First, ResNeXt101 is selected as the feature extractor to extract the features of each layer of the image. Second, MAM is used to generate a global semantic feature to guide the decoding process, and the global semantics are integrated into each layer feature of the decoder through upsampling, convolution, and element accumulation. In addition, the low-level features with more boundary information are generated after ERFM. Finally, features at all levels are sent into MLAM for effective aggregation of features, and the final significance map is generated through the cascade. Related contents will be described in detail in the following chapters.

3.1. Mixing Attention Module. Images are sent into the network and encoded to generate a series of features with different information. The features of the highest level have the strongest semantic representation ability, gradually fuse with the features of the lower level in the decoding process, and finally obtain the saliency map. However, the direct decoding and fusion of such semantic information will result in the loss of significant details. The reason is that different channels of high-level features and different spatial locations differently contribute to significance calculation. In particular, different channels may have different responses to the same object, and different spatial locations of the same channel may contain different objects. Inspired by the literature [15], this study designs the mixing attention module (MAM), and the module is divided into two parts, respectively, that is channel attention mechanism and spatial attention mechanism, which are used to capture different channels and the most significant part of the different space position, use the most significant semantic information, effectively enhance the characteristic of high rise, and get more robust global semantic characteristics. Figure 3 shows the detailed structure of the module.

3.1.1. Spatial Attention Mechanism. For the high-level features extracted from residual block 5, the width and height dimensions are first expanded into one-dimensional vectors and transposed to obtain the two-dimensional matrix $I \in \mathbb{R}^{\mathrm{BM}\times C}$, where C is the channel number of the feature, and B and M are the height and width. Then, through three parallel full-connection layers, M_v, M_z, and M_q, the dimension of the channel is reduced to obtain three matrices $V = IM_v$, $Z = IM_z$, and $Q = IM_q$, respectively. Then, $G = VZ^N$ is used to obtain the correlation matrix, where G_{xy} represents the inner product of the x th row in V and the y th row in Z, that is, the correlation of vectors at two different spatial positions. Each line of correlation matrix G is normalized by softmax function and constrained to (0, 1). Finally, the correlation matrix G is multiplied by Q, and the channel dimension is restored through G full-connection layer M_s, and the feature graph with enhanced spatial significance $I^S = GQM_s$ is obtained. The final feature expression is as follows:

$$I^S = \sigma\left(IM_v\left(IM_z\right)^N\right)IM_qM_s, \tag{1}$$

Among them, the $M_v, M_z, M_q \in \mathbb{R}^{C\times C/4}$, $M_s \in \mathbb{R}^{C/4\times C}$, and $\sigma(\cdot)$ are the softmax function.

3.1.2. Channel Attention Mechanism. The operation of the channel dimension is similar to the above. The features extracted from residual block 5 are first expanded into one dimension along with the width, then the high dimension, and then, transposed. The obtained $I \in \mathbb{R}^{BM\times C}$ passes through three full-connection layers and the outputs $V = IM_v$, $Z = IM_z$, and $Q = IM_q$. It is considered

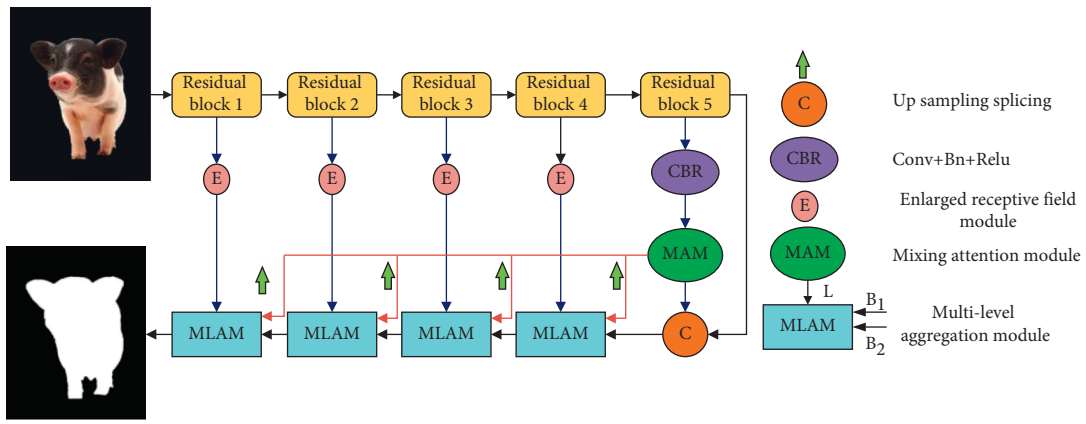

Figure 2: Network structure diagram.

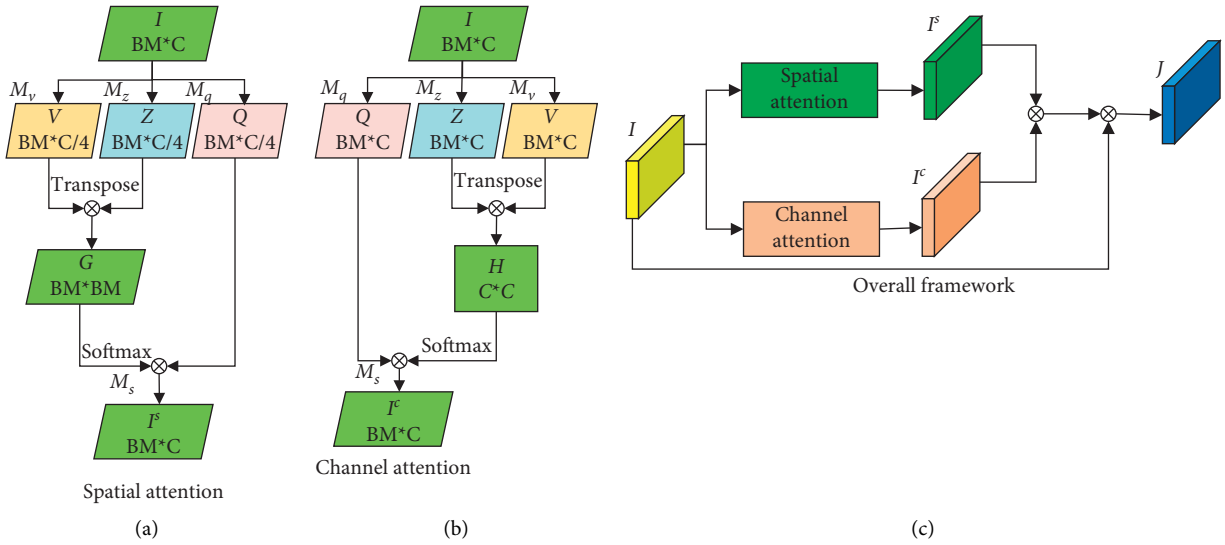

Figure 3: Mixing attention module. (a) Spatial attention, (b) channel attention, and (c) overall framework.

that dimensionality reduction will bring too much information loss. Therefore, this algorithm does not reduce the dimension of the channel. Then, the correlation matrix is obtained by $H = Z^N V$, where H_{xy} represents the inner product of the x th column in Z and the y th column in V, that is, the correlation of two different channel vectors. Similarly, each column of the correlation matrix H needs to be normalized to (0, 1) by the softmax function. Finally, by multiplying Q and H and passing through a fully connected layer M_s, the feature graph with enhanced spatial significance $I^C = QHM_s$ is obtained. The final feature expression is as follows:

$$I^C = IM_q\sigma\left(\left(IM_z\right)^N IM_v\right)M_s, \tag{2}$$

where $M_v, M_z, M_q, M_s \in \mathbb{R}^{C\times C}$. Finally, the output of the two branches is combined. Considering the influence of residual structure, this study adds the combined features and input I to generate the final feature graph $J \in \mathbb{R}^{BM\times C}$, which is formulated as follows:

$$J \in \mathbb{R}^{BM\times C}, \tag{3}$$

where + represents the addition of feature graphs at the element level. J is fed into the subsequent module after transposing and recovering the dimension expansion.

3.2. Enhanced Receptive Field Module. Although for the low-level features, the edges are very detailed. However, due to the limited number of downsampling, the receptive field is relatively limited and the global information cannot be captured. In the decoding process, if only the low-level features are simply used, although, the edge details are utilized. But the spatial details of features are not fully explored. Inspired by the literature [16], this study designs the enhanced receptive field module (ERFM) as shown in Figure 4. After low-level features pass through this module, the receptive field is enlarged and more spatial details are provided on the premise that edge details are not lost.

First, four parallel branches $(l_x, x = 1, 2, 3, 4)$ are designed for feature $W \in \mathbb{R}^{C\times B\times M}$, where l_1 adopts a

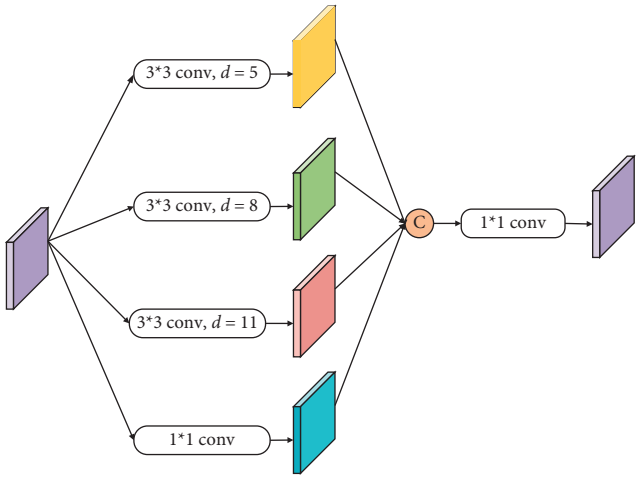

Figure 4: Enlarged receptive field module.

convolution kernel of 1×1, and the remaining three branches adopt a convolution kernel of 3×3. Different void rates are set for the three branches. Different voids are set according to the resolution of low-level features. For features with smaller resolution, smaller void rate is set, and for features with larger resolution, larger void rate is set. In this study, the maximum void rate is set as $D = (5, 8, 11)$, and it keeps shrinking with the narrowing of the feature map. Its specific setting will be explored in the ablation experiment. Then, the output of the four branches is spliced with channel dimension, and a 1×1 convolution is used to obtain the fused features.

3.3. Multilevel Aggregation Module. In the decoding process, it is crucial to make efficient use of the features of each layer. Previous related works only carried out simple splicing and fusion of high-level features and low-level features, and the results were very rough. Therefore, this study designs a multilevel aggregation module (MLAM). It can effectively aggregate features from different layers and different spatial scales. The input of this module is divided into three parts: semantic feature B_1 generated by MAM, low-level feature L enhanced by ERFM, and feature B_2 currently decoded. Figure 5 is a schematic of this module.

The whole polymerization process is divided into two stages. The first stage is the fusion of semantic features with the current decoding features. First, B_1 is convolved with two parallel 1×1 convolutions. After the splicing and fusion of the first branch and B_1 in the channel dimension, it is added with the results of the second branch to complete the first fusion, and the high-level feature B is obtained. Formulaic expression is as follows:

$$B = f_{\text{conv}}\left(f_{\text{cat}}\left(f_{\text{conv}}\left(B_1\right), B_2\right)\right) + f_{\text{conv}}\left(B_1\right), \tag{4}$$

$f_{\text{conv}}(\cdot)$ refers to the convolution operation, and $f_{\text{cat}}(\cdot)$ refers to the splicing operation of channels. The second stage is the aggregation of the high-level feature B obtained by the first-stage fusion and the low-level feature L enhanced by ERFM. This stage is divided into two parallel branches: bottom-up and top-down. Bottom-up is the aggregation of B to L, at which L remains unchanged. After an upsampling and a 1×1 convolution, B is spliced with L to obtain the aggregation graph $I^{b \longrightarrow l}$. Formulaic expression is as follows:

$$I^{b \longrightarrow l} = f_{\text{conv}}\left(f_{\text{cat}}\left(L, f_{\text{up}}\left(B\right)\right)\right). \tag{5}$$

Among them, $f_{\text{up}}(\cdot)$ refers to the upsampling operation. Top-down is the aggregation from L to B, in which B remains unchanged, and L first goes through a parallel pooling operation. Maximum pooling can extract the information with a larger response value in features, that is, the salient information contained in features. Average pooling can obtain global information on features. After such parallel pooling, feature L has stronger representational power and has the same spatial size as B. At this point, it is spliced with feature B in the channel dimension, and the fusion is completed by 1×1 convolution. Finally, upsampling is performed to obtain the final $I^{l \longrightarrow b}$. Formulaic expression is as follows:

$$I^{l \longrightarrow b} = f_{\text{up}}\left(f_{\text{conv}}\left(f_{\text{cat}}\left(B, f_{\text{avg}}\left(L\right) + f_{\text{max}}\left(L\right)\right)\right)\right), \tag{6}$$

$f_{\text{max}}(\cdot)$ and $f_{\text{avg}}(\cdot)$, respectively, represent maximum-pooling and average-pooling operations. Finally, the aggregation features obtained from the two branches are also aggregated. The final expression is as follows:

$$K = f_{\text{conv}}\left(f_{\text{cat}}\left(I^{l \longrightarrow b}, I^{b \longrightarrow l}\right)\right). \tag{7}$$

4. Experiment and Analysis

This part first introduces the experimental settings, including the experimental environment, datasets, comparison methods, and evaluation indicators. Then, the system verifies that the proposed algorithm can effectively improve the accuracy of image recognition through experimental comparison with a variety of comparison methods.

4.1. Experimental Settings. Experimental Environment. The experimental environment was four NVIDIA 1080TI GPUs, CUDA 8.0, and CUDNN 7.0. The batch size of experimental training was set to 32. In the experiment, ImageNet 2012 dataset was used for pretraining of VGG16. Without loss of generality, the learning rate of training in the initial state is set as 2×10^{-3}. At the 300th and 350th training periods, the learning rate was adjusted to 2×10^{-4} and 2×10^{-5}, respectively. At the 400th cycle, the training ended.

Dataset. The experimental datasets include PASCAL VOC [17]and MS COCO [18]. PASCAL VOC and MS COCO datasets contain 20 and 80 object classes, respectively. In the PASCAL VOC dataset, the training dataset is PASCAL VOC's trainval training dataset, and the test dataset is PASCAL VOC's test dataset. In the MS COCO dataset, the training dataset is trainval35 K. It contains 80,000 images, and the rest are a test dataset.

Contrast Method. The algorithm in this study adopts VGG16 as the backbone network. Therefore, VGG16 image recognition methods were used for comparison, including

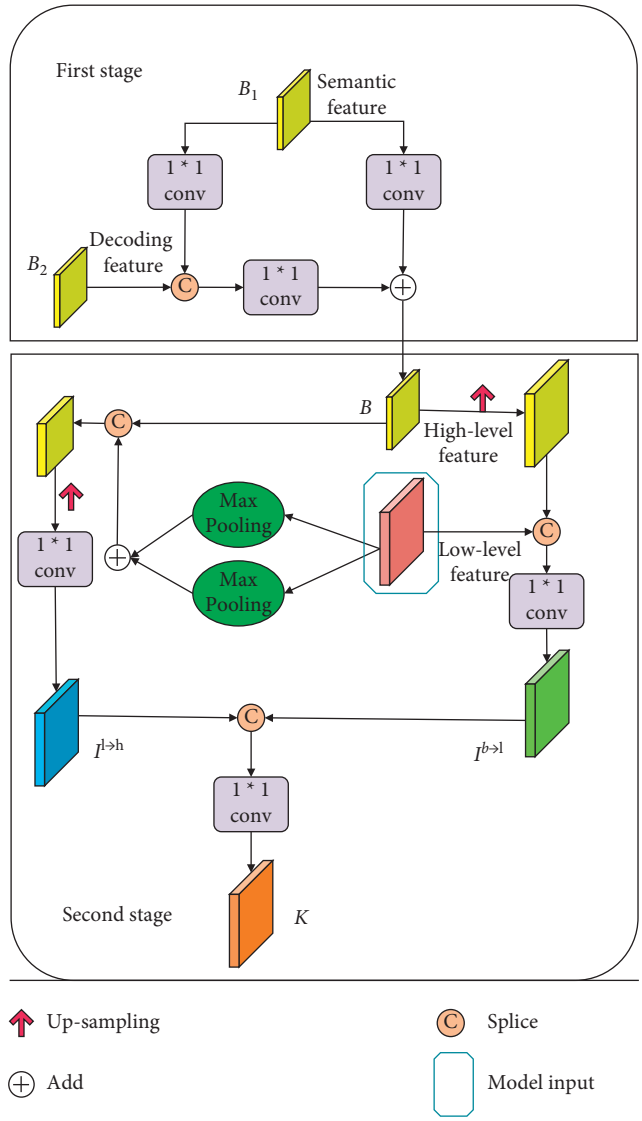

Figure 5: Multilevel fusion module.

literature [19], literature [20], literature [21], and literature [22].

Evaluation Indicators. In this study, the average accuracy of AP (average precision) and mAP (mean average precision) are selected as the core indexes of multiscale object recognition performance. AP and mAP are defined as follows:

$$\begin{aligned} \mathrm{GU} &= \frac{\sum_{r\in R} u(r)}{|R|}, \\ \mathrm{mAP} &= \sum_{x=1}^{X} \mathrm{AP}(x), \end{aligned} \tag{8}$$

where R is the set of recall rate. U (r) is the accuracy when the recall rate is r. X is the total number of categories. AP(x) is the average accuracy of classification x.

4.2. Experimental Results. Table 1 shows the average accuracy of experiments on 20 classes of objects in the PASCAL VOC dataset. The average accuracy of the proposed algorithm on the PASCAL VOC dataset is 80.4%, which is the best among all the algorithms. The mean accuracies of literature [22], literature [21], literature [20], and literature [19] were 80.9%, 79.4%, 76.8%, and 74.4%, respectively. Obviously, the algorithm in this study has achieved the best accuracy in image recognition.

In order to further verify the accuracy of the proposed algorithm for image recognition, this part is further verified on MS COCO dataset. In the experiment, literature [19] is a two-step image recognition method, while other methods are single-step image recognition methods. In image recognition, the finer the image, the larger the size of the original input image; and the more information in the image, the better is the image recognition effect. In the single-step comparison method,

Table 1: Experimental results of PASCAL VOC dataset.

Methods	Literature [19]	Literature [20]	Literature [21]	Literature [22]	Proposed
mAP (%)	74.4	76.8	79.4	80.9	**82.6**

Bold value represents the ideal value.

both the latest literature [21] and literature [22], and the algorithm in this study set the image input size as 512×512 to compare the experimental results.

Table 2 shows the experimental results on the MS COCO dataset. Frame per second (FPS) refers to the number of detected images per second. Obviously, under the same input conditions, the proposed algorithm has a lower FPS than the literature [21] and literature [22] algorithms. In other words, the operation efficiency of the proposed algorithm is better than that of the literature [21] and literature [22] algorithms. Experiments also verify the influence of different IoU (intersection over union) on image recognition accuracy. In the experiment, IoU was set to 0.5, 0.75, and 0.95, respectively. It is not hard to see that with the increase in IoU value, the average accuracy of all algorithms decreases. However, in three different IoU experiments, the proposed algorithm achieves the best average accuracy of multiscale image recognition. Literature [22] is a better algorithm for image recognition. Therefore, this part focuses on comparing the algorithm proposed in this study with the literature [22] algorithm. Obviously, when IoU is 0.5, 0.75, and 0.95, the average accuracy of this algorithm is 58.4%, 39.1%, and 33.8%, respectively, which are 3.9%, 3.6%, and 0.8% higher than the literature [22] algorithm. Finally, the experiment verifies the recognition accuracy of objects with different scales when IoU = 0.75. As can be seen from Table 2, the recognition accuracy of the algorithm in this study for small-scale, medium-scale, and large-scale objects is 16.6%, 37.8%, and 45.2% respectively, which is 0.4%, 1.4%, and 0.8% higher than that of the literature [22] algorithm. Experimental results show that the proposed algorithm can effectively improve the recognition accuracy of multiscale objects.

The MR (miss rate) indicates the miss rate commonly used in the target detection field. FPPI (false positives per image) refers to the error detection rate per frame (the ratio of the number of negative samples predicted by the model as positive samples to all samples). The lower the MR-FPPI curve is, the better is the performance of the graph perception algorithm on the test set. Figures 6 and 7 show the MR-FPPI changes in the proposed algorithm and various comparison algorithms on the PASCAL VOC dataset and the MS COCO dataset, respectively. As can be seen from the figure, the curve of the proposed algorithm is the lowest on PASCAL VOC datasets and MS COCO datasets, and rapidly decreases, achieving the best detection performance.

This section also explores the influence of iteration times on model's learning efficiency. As shown in Figure 8, the training accuracy of the algorithm in this study tends to be stable after about 10 cycles of iteration, and the model begins to converge.

Table 2: Experimental results of MS COCO dataset (bold is the best result).

Methods	Input size	FPS	Average accuracy (IoU)			Average accuracy (scale)		
			0.5	0.75	0.95	S	M	L
Literature [19]	1000×600	7	42.7	-	21.9	—	—	—
Literature [20]	384×384	15	49.5	27.1	27.4	—	—	—
Literature [21]	512×512	22	48.5	30.3	28.8	10.9	31.8	43.5
Literature [22]	512×512	22.3	54.5	35.5	33	16.3	36.3	44.3
Proposed	512×512	**20.7**	**58.4**	**39.1**	**33.8**	**16.7**	**37.7**	**45.1**

Bold values represent the ideal values.

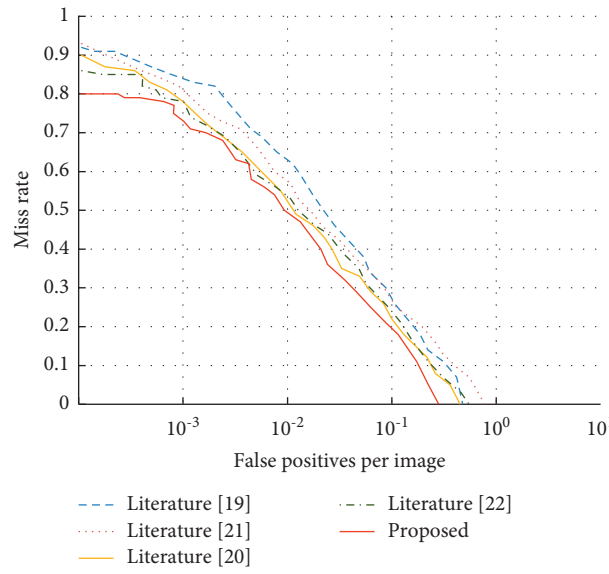

Figure 6: The MR-FPPI result of the PASCAL VOC dataset.

The graphic perception system designed in this study can also map the received binary edge image data to the electrode array so that the blind person can generate the corresponding electrical stimulation. Figure 9 shows the schematic diagram of electrical stimulation corresponding to a two-dimensional electrode array when the camera shoots several aircraft. The solid circle represents electric stimulation, and the hollow circle represents no electric stimulation. Blind people can perceive what they are "seeing" as an object, for example, an airplane, through electrical stimulation of the skin on their abdomen.

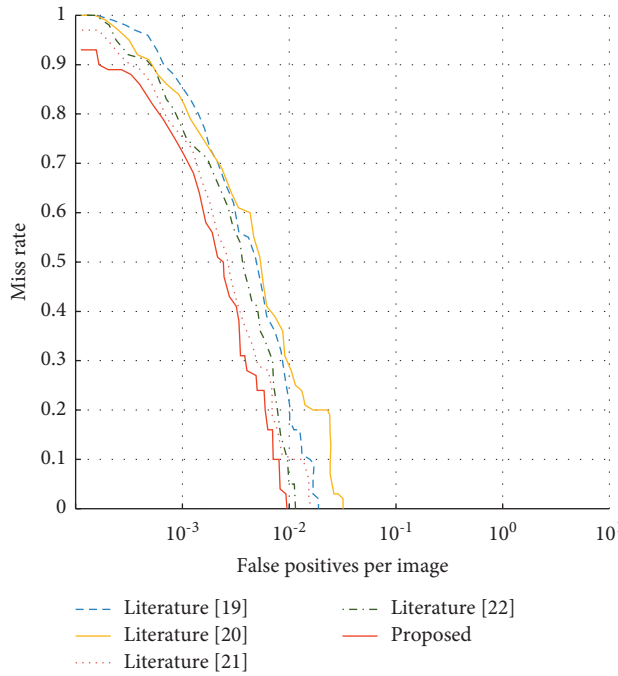

Figure 7: The MR-FPPI result of the MS COCO dataset.

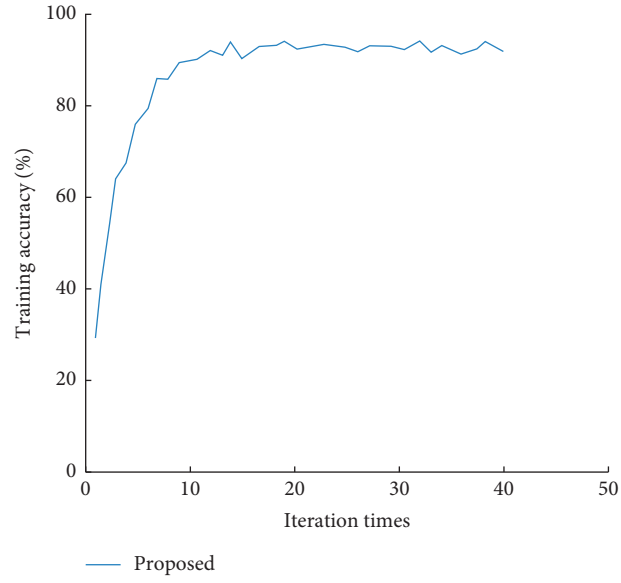

Figure 8: Training accuracy curve.

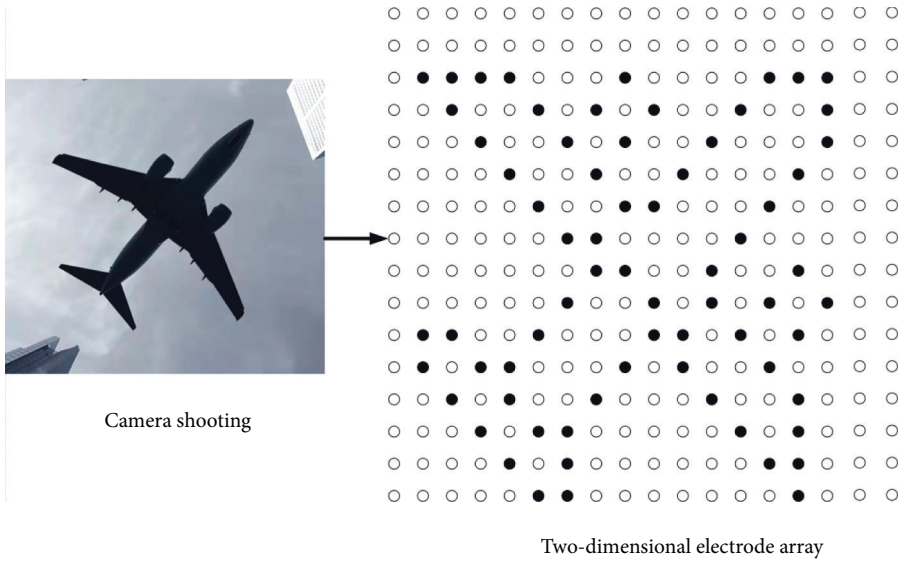

Figure 9: Schematic diagram of two-dimensional electrode array electrical stimulation.

5. Conclusions

Information is the basis for people to acquire knowledge and communicate with each other. However, for people with visual disabilities, because of visual impairment, other perceptual abilities, such as hearing, touch, smell, and taste perception, have become the main channels for them to obtain information and explore their surrounding environment. Image object recognition is a basic task in the field of computer vision. However, existing methods cannot improve the accuracy rate of object recognition for visually impaired groups. In order to solve this problem, this study proposes a graphic perception system that improves the attention mechanism. The system consists of three parts: mixing attention module, enlarged receptive field module, and multilevel aggregation module. First, the module of increasing the receptive field is used to process the low-level features extracted from the feature extraction network so that it can increase the receptive field while retaining the original edge details, so as to obtain more abundant graphics and image information. Then, the last layer of the feature extraction network is processed by the mixing attention module to enhance its representational power and serve as semantic guidance in the decoding process to guide the feature aggregation. Finally, the multilevel aggregation module can effectively aggregate the features from different levels to obtain the perceptive object graphics. Compared with other algorithms on PASCAL VOC and MS-COCO data, the experimental results show that the proposed method can effectively improve the accuracy of graphic object recognition.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the Natural Science Foundation of Hunan Province in 2021: "Research on Tactile Graphic Design and Application for Visually Impaired Population in the Era of Graphic-Based Reading," project number: 2021JJ60022.

References

[1] M. Hu, Y. Chen, and G. Zhai, "An overview of assistive devices for blind and visually impaired groups," *International Journal of Robotics and Automation*, vol. 34, no. 5, pp. 580–598, 2019.

[2] S. S. Senjam, "Impact of COVID-19 pandemic on people living with visual disability," *Indian Journal of Ophthalmology*, vol. 68, no. 7, p. 1367, 2020.

[3] A. Panotopoulou, X. Zhang, T. Qiu, and E. Whiting, "Tactile line drawings for improved shape understanding in blind and visually impaired users," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 1–89, 2020, 89.

[4] E. R. Hofmann, C. Davidson, H. Chen et al., "Blind spot: a braille patterned Novel multiplex lateral flow immunoassay sensor array for the detection of biothreat agents," *ACS Omega*, vol. 6, no. 35, pp. 22700–22708, 2021.

[5] D. Prescher, J. Bornschein, W. Köhlmann, and G. Weber, "Touching graphical applications: bimanual tactile interaction on the HyperBraille pin-matrix display," *Universal Access in the Information Society*, vol. 17, no. 2, pp. 391–409, 2018.

[6] D. Karastoyanov, N. Stoimenov, and S. Gyoshev, "Methods and means for education of people with visual impairments," *IFAC-PapersOnLine*, vol. 52, no. 25, pp. 539–542, 2019.

[7] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Virtual-blind-road following-based wearable navigation device for blind people," *IEEE Transactions on Consumer Electronics*, vol. 64, no. 1, pp. 136–143, 2018.

[8] S. Mun, S. Yun, S. Nam et al., "Electro-active polymer based soft tactile interface for wearable devices," *IEEE transactions on haptics*, vol. 11, no. 1, pp. 15–21, 2018.

[9] S. Real and A. Araujo, "Navigation systems for the blind and visually impaired: past work, challenges, and open problems," *Sensors*, vol. 19, no. 15, p. 3404, 2019.

[10] A. van Leendert, M. Doorman, P. Drijvers, J. Pel, and J. van der Steen, "An exploratory study of reading mathematical expressions by braille readers," *Journal of Visual Impairment & Blindness*, vol. 113, no. 1, pp. 68–80, 2019.

[11] E. Devile and E. Kastenholz, "Accessible tourism experiences: the voice of people with visual disabilities," *Journal of Policy Research in Tourism, Leisure and Events*, vol. 10, no. 3, pp. 265–285, 2018.

[12] S. Chen, K. Jiang, Z. Lou, D. Chen, and G Shen, "Recent developments in graphene-based tactile sensors and E-skins," *Advanced Materials Technologies*, vol. 3, no. 2, Article ID 1700248, 2018.

[13] T. Graven, I. Emsley, N. Bird, and S. Griffiths, "Improved access to museum collections without vision: how museum visitors with very low or no vision perceive and process tactile–auditory pictures," *British Journal of Visual Impairment*, vol. 38, no. 1, pp. 79–103, 2020.

[14] L. Cavazos Quero, J. Iranzo Bartolomé, and J. Cho, "Accessible visual artworks for blind and visually impaired people: comparing a multimodal approach with tactile graphics," *Electronics*, vol. 10, no. 3, p. 297, 2021.

[15] B. Leporini, V. Rossetti, F. Furfari, S. Pelagatti, and A. Quarta, "Design guidelines for an interactive 3D model as a supporting tool for exploring a cultural site by visually impaired and sighted people," *ACM Transactions on Accessible Computing*, vol. 13, no. 3, pp. 1–39, 2020.

[16] S. Fernando and J. Ohene-Djan, "An empirical evaluation of a graphics creation technique for blind and visually impaired individuals," *British Journal of Visual Impairment*, vol. 39, no. 3, pp. 191–213, 2021.

[17] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Scientific Reports*, vol. 10, no. 1, pp. 11307–11313, 2020.

[18] X. Li, C. Xu, X. Wang et al., "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2347–2360, 2019.

[19] R. Meng, S. G. Rice, and J. Wang, "A fusion steganographic algorithm based on faster R-CNN," *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1–16, 2018.

[20] O. Badmos, A. Kopp, T. Bernthaler, and G. Schneider, "Image-based defect detection in lithium-ion battery electrode using convolutional neural networks," *Journal of Intelligent Manufacturing*, vol. 31, no. 4, pp. 885–897, 2020.

[21] S. Singhal, V. Passricha, P. Sharma, and R. K. Aggarwal, "Multi-level region-of-interest CNNs for end to end speech recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 11, pp. 4615–4624, 2019.

[22] C. Sun, Y. Ai, S. Wang, and W. Zhang, "Dense-RefineDet for traffic sign detection and classification," *Sensors*, vol. 20, no. 22, p. 6570, 2020.