Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 3, 2024

Research on Intelligent System of Multimodal Deep Learning in Image Recognition

Ting Xu[1, a], Iris Li [2, b], Qishi Zhan [3, c], Yuxiang Hu [4, d], Haowei Yang [5, e]

[1] University of Massachusetts, Boston, USA
[2] New York University, USA
[3] Marquette University, USA
[4] Johns Hopkins University, USA
[5] University of Houston, USA

[a]ting.xu001@umb.edu; [b]irisepode@gmail.com; [c]qishizhan7@gmail.com; [d]yhu102@jhu.edu; [e]yanghaowei09@gmail.com

Abstract: In this paper, a multi-scale image estimation method based on wavelet transform is proposed, which can effectively remove motion features from multiple videos. Then the autoencoder with sparsity limit is used to adjust the input signal to achieve effective compression. The effective features are extracted and the optimal unique vector is learned. The improved convolutional neural network is used to recognize weak moving objects. Experiments show that the algorithm can achieve high accuracy without large-scale learning samples, and the highest recognition rate is 99.36%. This algorithm has a great improvement over conventional algorithm.

Keywords: Single Frame Image Denoising; Object Segmentation; Sparse Representation; Deep Neural Network.

1. Introduction

Image sequence refers to a series of images obtained consecutively from the target at different times and in different directions, which is a kind of multi-modal image. Generally, in complex continuous images, weak moving objects usually appear only a single or a few pixels, and are susceptible to strong waves, and the signal-to-noise ratio is relatively poor, so how to efficiently and accurately identify them has become a research focus.

Deep learning is a concept that evolved from studying neural networks. It involves breaking down the learning structure into hidden and multi-layer perceptron's. The key idea is to imitate the brain's processes to study and understand data, and then interpret the findings in a way that mirrors how the brain works. Currently, there's extensive research focused on detecting weak moving objects in video sequences. One approach to this uses wavelet signal transformation to defocus an image, minimizing the wavelet coefficient values [1]. After that, other noise-related coefficients are removed, followed by dynamic object clustering. This technique is quick but has limitations when it comes to identifying similar targets accurately. Another approach uses time-space non-local similarity to segment weak moving objects [2]. This method involves matching similar features in multiple images to create a segmented time-space field. Though it has a high recognition rate, it struggles with distinguishing similar targets.

In recent years, with the development of deep learning technology, it has made important breakthroughs in many fields [3][4]. For example, convolutional neural networks (CNNs) perform well with medical recognition but encounter limitations in computing power, network depth, and optimization algorithms when applied to other types of recognition [5][6]. Another study that used an 8-layer deep CNN for image recognition reported good results, leading to further applications in image processing [7]. Another study indicates that a 2D convolutional neural network can detect the time characteristics, but due to the constraints of the network itself, it is only suitable for short-time video, and cannot meet the demand of massive data[8]. The LRCN algorithm proposed in the literature [9] overlaps the results of temporal network and CNN, thus taking into account the characteristics of both time and time. However, the uncertain information of other modes makes the method cannot be applied well. The dual-stream algorithm proposed improves image identification efficiency by adding optical flow to conventional RGB images, but does not effectively mine image information containing features such as depth and skeleton in images [10]. In addition, it is more difficult to detect real-time images because of the increase in optical flow computation. At present, a lot of research work is based on the feature vocabulary package to achieve the extraction and expression of spatiotemporal interest points. However, due to the impact of shooting Angle, shaking, complex background, and other factors, the algorithm is faced with great challenges. Studies have shown that Hamiltonian corner points are extended to 3D scenes, and traditional feature representation methods such as SIFT and HOG are extended to SIFT 3D and HOG-3D [11]. Although several feature extraction algorithms based on manual design have obtained good results, their performance on large samples is not ideal. The deep convolutional neural network is a kind of neural network with a multi-level structure, which often convolves or centralizes the variation in order to obtain the characteristics of the input. However, due to the constraints of data sources, its application in image sequences is more complex, usually only applicable to the image of a single image.

Recurrent neural networks (RNN) are very different from traditional forward neural networks [12]. In this method, the inner information of neurons is collected by forming a directed ring of neurons, so that the time series can be sensed. At present, recurrent neural networks have been widely used in speech recognition. In the past, the hidden Markov model was mainly used, but with the popularization of deep network, it is gradually replaced by Markov model. In the recurrent

neural network, the basic delay backward transfer algorithm is used to correct the weights and solve the network errors. However, classical recurrent neural networks do not work well with historical data. Some scholars proposed short-term memory neural network for the first time in 1997, and added new functions such as input port and amnesia gate, so that it can selectively remember the key information at a certain point in time.

2. Design of intelligent image recognition system.

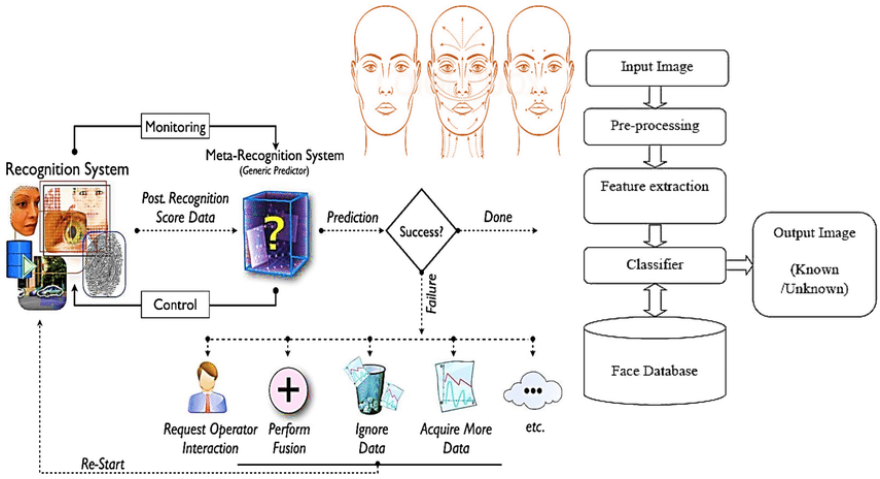

Fig. 1 Structure of image recognition system

The Unet network is used to realize the efficient connection of the underlying features, the efficient connection between the underlying features and the underlying features, and the efficient fusion of the underlying features [13]. In order to solve the problem of difficulty in obtaining image features due to the large Angle difference between moving end and complex scene, an image classification method based on OpenCV is proposed to realize automatic face recognition. the structure of the image recognition system is shown in Figure 1 (image cited in LAMSTAR: For IoT-based face recognition system to manage the safety factor in smart cities)

2.1. U-net network image semantic segmentation

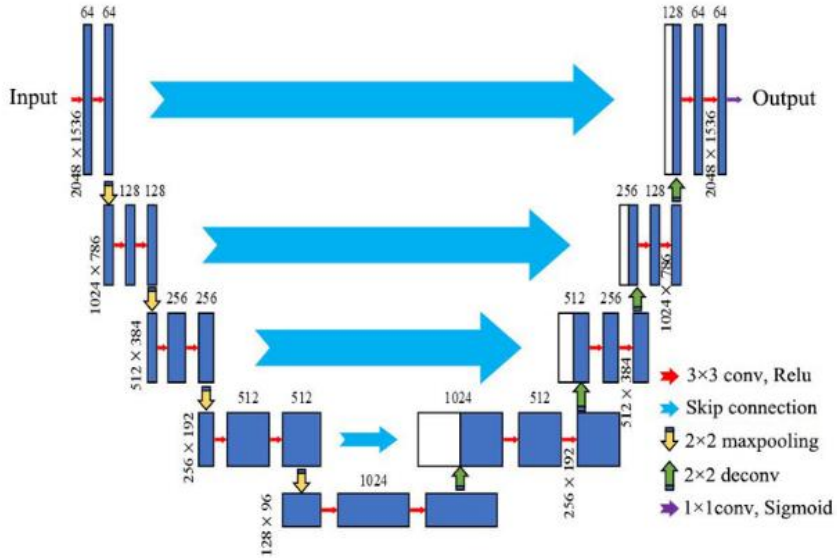

Fig. 2 U-net neural network structure

In practice, photos taken on Android phones are affected by factors such as shooting Angle, lighting and background, making it difficult to accurately segment different areas. A method of semantic segmentation of Unet using deep neural networks is proposed. U-shaped network can be divided into left and right regions on the premise that the size of the original image is equivalent. The left side is the compressed channel, and the right side is the expanded channel, also known as up-sampling and down-sampling. Features are obtained through down-sampling, and the number of image channels will be doubled during down-sampling. The feature map dimension of the image will also be reduced to half of the original. The adaptability to nonlinear modeling is realized by the improved ReLU excitation function. Through upper sampling, the dimension of the feature graph is reduced to the same degree as the input graph, so that when passing the upper sampling layer, the dimension of the feature graph is doubled, and the number of channels is reduced by half. Finally, the original data is restored to the original size. In this algorithm, the compressed channel is composed of 4 scaling elements, each of which is composed of 2 convolution and 1 pool, and is up-sampled after the end of down sampling. Among them, the expansion channel is composed of 4 expansion cells, and each expansion cell is composed of upper sampling layer, convolution and 1x1 convolution module. The network has a U-shape, so it is widely used in semantic

segmentation of medical images, traffic signs and other fields. The input image size of the network is 512x512, and the network structure is shown in Figure 2.

2.2. Convolutional neural network character recognition

Convolution is a common operation in image processing. This algorithm is not only suitable for image denoising, enhancement, edge detection, prediction and other fields [14], but also suitable for image feature extraction [15]. The operation steps of two-dimensional convolution are given in Figure 3.

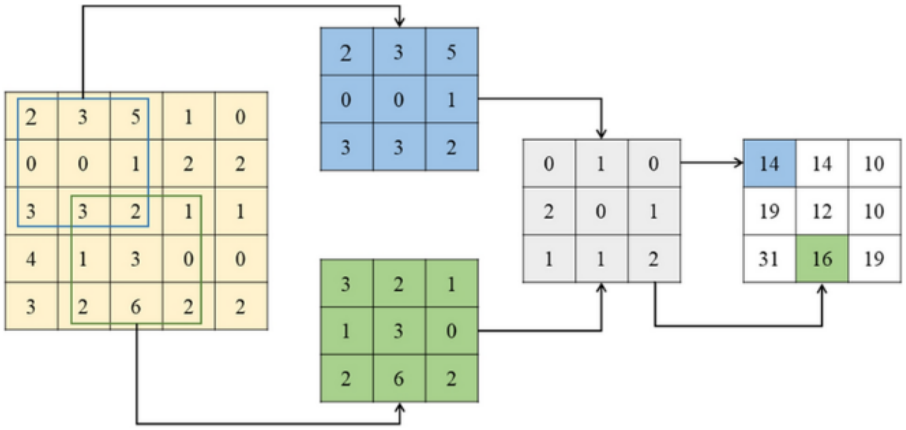

Fig.3 Convolution calculation process

The input of the convolutional network is generally a multi-channel plane model with dimensions of length × width × number of channels, that is, a multi-channel 2D characteristic graph. An important parameter of the convolutional neural network is the convolution kernel function. Multiple convolution functions are used to convolve images from left to right and from top to bottom to generate feature images of multiple images. By adding convolution functions, the complex and abstract internal structure of the image can be extracted, and then multiple convolution calculations can finally realize the abstract expression of the image at multiple levels. Convolutional neural network architecture is shown in Figure 4 (the picture is quoted in How to draw Deep learning network architecture diagrams?).

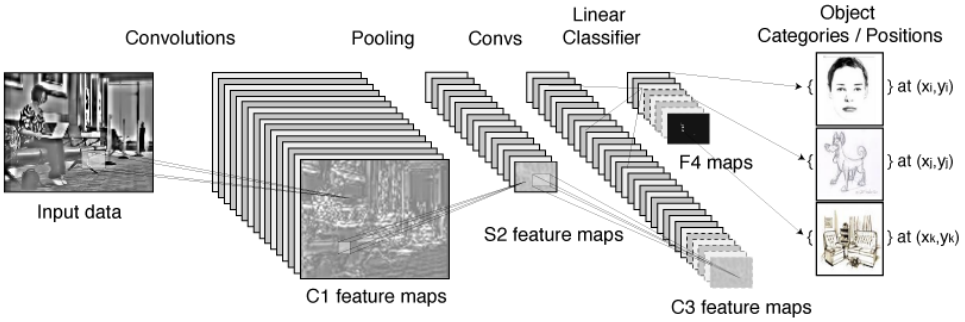

Fig. 4 Convolutional neural network structure diagram

3. Dim Unit target recognition algorithm based on CNN multi-modal learning

Each feature mapping layer is regarded as a two-dimensional space with the same weight, and each feature is mapped to the two-dimensional space with the same weight. In this project, the sigmoid function is used as the excitation function, and a feature mapping of location invariants is established. Through the interrelation of each layer and operation layer of the convolutional neural network, the local mean is obtained, and then the secondary extraction is carried out. Each convolution layer C_1 performs linear $K^l \times K^l$ filtering on the input plane $z_{1,\ \cdots,\ N^{l-1}}^{l-1}$ of the $N^{y^{-1}}$ layer. The value of N^l at $(i,\ j)$ in the p^{th} plane is calculated as follows

$$z_p^l(i,\ j) = b_p^l + \sum_q \sum_{s=1}^{K^l} \sum_{i=1}^{K^l} w_{p,\ q,\ s,\ t}^l z_p^{l-1}(i-1+s,\ j-1+t) \quad (1)$$

The bias b_p^l and filter weight $w_{p,\ q,\ s,\ t}^l$ are trained by backpropagation algorithm. The output plane is $D^{l-1} \times D^{l-1}$, where $D^l = D^{l-1} - K^l + 1$. The subsampling layer S_l is smoothed by $K^l \times K^l$ on each access plane

$$z_p^l(i,\ j) = b_p + w_p \sum_{s=1}^{K^l} \sum_{t=1}^{K^l} z_p^{l-1}(i-1+s,\ j-1+t) \quad (2)$$

"Softmax" layer is introduced to explain these vectors, and the calculation formula is as follows

$$\tilde{P}_p = \frac{\exp\left(z_p^{l-1}\right)}{\sum_q \exp\left(z_p^{l-1}\right)} \quad (3)$$

$$L(\theta) = -\sum_{n-1}^{N} \lg P_\theta(y_n \mid x_n) = -\sum_{n-1}^{N} \lg \tilde{P}_{\theta,\ y_n}(x_n) \quad (4)$$

After optimization, the parameter θ is trained by the stochastic gradient descent algorithm, and the gradient $\partial L(\theta) / \partial \theta$ of the random sample $(x,\ y)$ is calculated, and then updated:

$$\theta - \lambda \frac{\partial L(\theta,\ x,\ y)}{\partial \theta} \to \theta \quad (5)$$

In the process of dim dim target recognition, this method uses the fixed features of continuous and non-continuous frames to optimize the recognition method, so as to ensure the

high accuracy of the recognition results.

4. System Simulation

Through simulation experiments, the proposed algorithm is studied experimentally. Using small-scale Imagenet image data as the research object, MATLAB simulation software was used to conduct experiments, and a 1.0Gb image was selected as the training image. Through the identification method of a multi-modal deep network, the weak object identification method based on the multi-modal deep network is studied. The reasonable values of each parameter are obtained by testing the sampled images. Firstly, the corresponding convolution features will be extracted from the existing image, and then the specific values will be determined according to the existing image structure and the sampled image. The effect of these methods on the accuracy of images with different characteristics is shown in Figure 5.

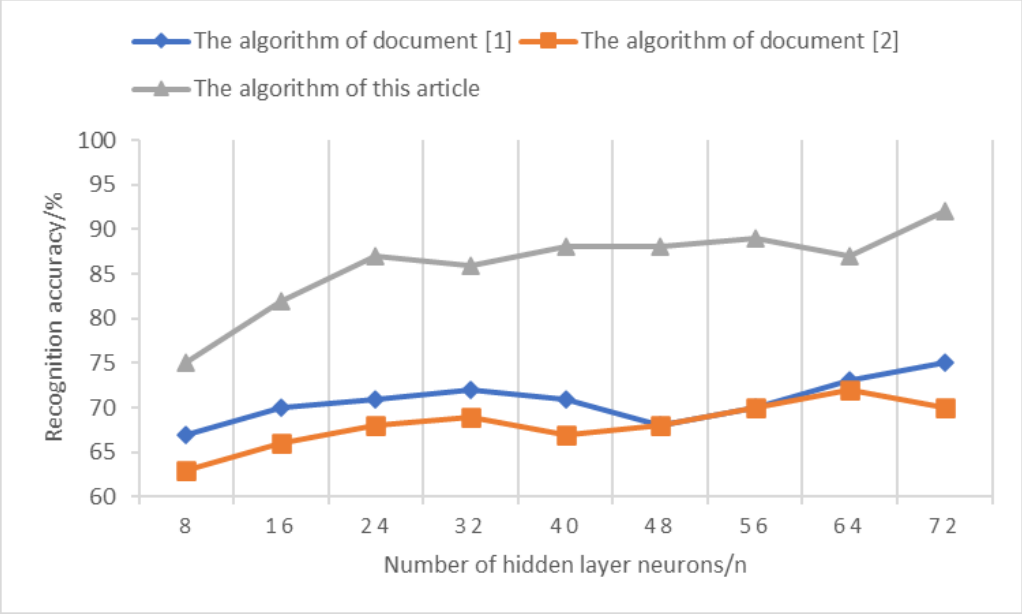

Fig. 5 Effect of recurrent neural networks and the number of hidden neurons on accuracy

When the number increases, the recognition accuracy of this method is gradually improved, and when the number is 72, the accuracy of this method is the highest. Therefore, when the algorithm is used to classify the weak objects in an image sequence, the number is set to 72, so that the most accurate object recognition rate can be obtained. However, the accuracy of existing object recognition algorithms based on text comparison is not as good as that of the algorithm proposed in this paper, and its accuracy also fluctuates greatly.

5. Conclusion

This project intends to carry out research on weak moving target identification in multi-mode deep neural networks. Through image preprocessing and sparse self-coding and decoding, an improved convolutional neural network is used to recognize weak-moving objects. Through experiments, the new algorithm proposed in this paper can realize high-precision image recognition without relying on large-scale sample data, so as to achieve the purpose of high efficiency and accuracy.

References

[1] Wang, X. S., Turner, J. D., & Mann, B. P. (2021). Constrained attractor selection using deep reinforcement learning. Journal of Vibration and Control, 27(5-6), 502-514.

[2] Liu, Z., Yang, Y., Pan, Z., Sharma, A., Hasan, A., Ding, C., ... & Geng, T. (2023, July). Ising-cf: A pathbreaking collaborative filtering method through efficient ising machine learning. In 2023 60th ACM/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.

[3] Zi, Y., Wang, Q., Gao, Z., Cheng, X., & Mei, T. (2024). Research on the Application of Deep Learning in Medical Image Segmentation and 3D Reconstruction. Academic Journal of Science and Technology, 10(2), 8-12.

[4] Yan, C., Qiu, Y., Zhu, Y. (2021). Predict Oil Production with LSTM Neural Network. In: Liu, Q., Liu, X., Li, L., Zhou, H., Zhao, HH. (eds) Proceedings of the 9th International Conference on Computer Engineering and Networks . Advances in Intelligent Systems and Computing, vol 1143. Springer, Singapore. https://doi.org/10.1007/978-981-15-3753-0_34.

[5] Xin Chen , Yuxiang Hu, Ting Xu, Haowei Yang, Tong Wu. (2024). Advancements in AI for Oncology: Developing an Enhanced YOLOv5-based Cancer Cell Detection System. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),75-80, doi:10.55524/ijircst.2024.12.2.13.

[6] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024). Survival prediction across diverse cancer types using neural networks. doi:10.48550/ARXIV.2404.08713

[7] Li, S., Kou, P., Ma, M., Yang, H., Huang, S., & Yang, Z. (2024). Application of Semi-supervised Learning in Image Classification: Research on Fusion of Labeled and Unlabeled Data. IEEE Access.

[8] Yao, J., Wu, T., & Zhang, X. (2023). Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn. arXiv preprint arXiv:2308.08333.

[9] Yulu Gong , Haoxin Zhang, Ruilin Xu, Zhou Yu, Jingbo Zhang. (2024). Innovative Deep Learning Methods for Precancerous Lesion Detection. International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12(2),81-86, doi:10.55524/ijircst.2024.12.2.14.

[10] Xiao, M., Li, Y., Yan, X., Gao, M., & Wang, W. (2024). Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. doi:10.48550/ARXIV.2404.08279

[11] Guo, A., Hao, Y., Wu, C., Haghi, P., Pan, Z., Si, M., ... & Geng, T. (2023, June). Software-hardware co-design of heterogeneous SmartNIC system for recommendation models inference and training. In Proceedings of the 37th International Conference on Supercomputing (pp. 336-347).

[12] Hu, Z., Li, J., Pan, Z., Zhou, S., Yang, L., Ding, C., ... & Jiang, W. (2022, October). On the design of quantum graph convolutional neural network in the nisq-era and beyond. In 2022 IEEE 40th International Conference on Computer Design (ICCD) (pp. 290-297). IEEE.

[13] Wang, X. S., & Mann, B. P. (2020). Attractor Selection in Nonlinear Energy Harvesting Using Deep Reinforcement Learning. arXiv preprint arXiv:2010.01255.

[14] Dai, W., Tao, J., Yan, X., Feng, Z., & Chen, J. (2023, November). Addressing Unintended Bias in Toxicity Detection: An LSTM and Attention-Based Approach. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 375-379). IEEE.

[15] Liu, Y., Yang, H., & Wu, C. (2023). Unveiling patterns: A study on semi-supervised classification of strip surface defects. IEEE Access, 11, 119933-119946.