Appl. Comput. Math., V.22, N.3, Special Issue, 2023, pp.355-371
DOI: 10.30546/1683-6154.22.3.2023.355

DEEP LEARNING DRIVEN SMART EDUCATION: TEACHING MANAGEMENT MECHANISM BASED ON METAVERSE AND BIG DATA ANALYSIS

HONG ZHANG[1], YANAN SUN[1,*], ADAM SLOWIK[2], LINGLING ZHANG[1], JIANHUI LV[3]

Abstract. With the application of the metaverse and extended reality, education in the era of Artificial Intelligence (AI) will be greatly enhanced. However, how to apply this new technology to the smart education management mechanism is still a big challenge. In this paper, we use intelligent technology to empower education and propose a new method of teaching management and attendance. Specifically, we combine intelligent self-powered sensors and deep learning methods to propose a face recognition model with improved multi-level feature fusion. First, we employ the sparse coding algorithm to extract primary low-level features of face images. Second, we utilize a convolutional neural network model to learn high-level features. At the same time, we take intelligent self-powered sensors to collect data from remote scenarios. Finally, the face is recognized and classified based on the extracted feature knowledge vector. The experimental results on two datasets show that our method has a good recognition performance.

Keywords: Deep Learning, Self-Powered Sensor, Smart Education, Metaverse, Big Data, Convolutional Neural Network.

Mathematics Subject Classification: 68T07, 68T09, 68U07, 97P80.

1. Introduction

The metaverse is not an isolated technology, and it is one of many technologies in the ecosystem, such as cloud computing [34], high-speed networks, robotics, 3D printing, big data, artificial intelligence, nanotechnology, and the Internet of Things [1, 12, 26]. These technologies mark the transformation of the era of Industry 4.0 [23], closely linked to the Fourth Industrial Revolution, enabling automated intelligent manufacturing processes and devices that will have an impact on jobs and the economy. The new generation of learners and teachers need to be equipped with appropriate abilities, and the existing workforce needs to carry out large-scale reskilling and upskilling under the framework of lifelong learning. With the application of the metaverse and Extended Reality (XR) [10], education in the era of Industry 4.0 will be greatly enhanced. Related studies believe that the metaverse is a 3D virtual world, in which people and people are related to each other and have their own digital identities, and maybe everything can be created in the 3D virtual world [42]. The metaverse is the integration of the virtual world and the real world with the support of relevant technologies [15]. The metaverse is a digital world, which can realize the seamless connection between virtual and real, and the value generated can surpass the real world [14]. These studies define the meta-universe from a technical perspective, ignoring the value of human beings. The third wave of computing technology innovation represented by personal computers, the Internet and mobile devices has brought drastic changes to interactive communication, teaching and learning, business exchanges, and so on, and profoundly changed

[1]Beihua University, Jilin 132013, China, e-mail: zhanghong@beihua.edu.cn, zhanglingling@beihua.edu.cn
[2]Koszalin University of Technology, Koszalin, Poland, e-mail: adam.slowik@tu.koszalin.pl
[3]Pengcheng Laboratory, Shenzhen 518057, China, e-mail: lvjh@pcl.ac.cn
*Corresponding author: e-mail: sunyanan@beihua.edu.cn
Manuscript received 20 March 2023.

people's life. The fourth wave brought many new technologies, such as XR. XR is an umbrella term that includes a range of technologies such as virtual reality (VR), augmented reality (AR), mixed reality (MR) [16], and electronic digital environments where information is projected and people can observe and act in fully or partially digitized environments. People can better understand the above technology through the Reality-Virtuality Continuum [28]. People can project information into the physical environment through handheld devices or headsets. A fully virtual environment, also known as virtual reality, is an environment that is completely from of the physical world.

The metaverse is a perceptive social environment that integrates physical reality and digital virtuality [29]. It is an Internet composed of a socialized and networked immersive environment, enabling users to communicate in real time. If you want to explain the Metaverse to non-academics, you can also say that the metaverse is a kind of Post Reality or virtual universe, because the English word "metaverse" consists of two words: "meta" and "universe". Since it is a network in three-dimensional space, we can also call the metaverse 3D Internet or Internet of Place. The World Economic Forum has proposed seven important themes for the metaverse: Digital Teleportation, Technological barriers, Immersive Storytelling, The Gamification of Life, Health Benefits and Risks, Immersive Media and Art, and New Ways to Make, Do and Buy. Tony Parisi's Seven Laws of the metaverse [17] outline the characteristics of the metaverse. The metaverse is a manifestation of the Internet, which is an open, cross-hardware network. It's owned by everyone, not by an entity, and it's not controlled by a company. Of course, the interior of the metaverse is similar to the interior of the Internet in that it is sealable, allowing people to run companies, conduct academic activities and do other things in closed virtual spaces.

The upsurge of the metaverse has brought great impact on education. Researchers at home and abroad begin to attach importance to the deep integration of the metaverse and the educational ecosystem. The educational ecosystem consists of family education, school education and social education. Many scholars focus on school education and family education in the metaverse, but ignore social education, which is the only way to improve citizens' quality and happiness. Social education metaverse is not only a technical problem, but also a problem of social education concept and organizational reform, and more attention should be paid to human nature. The social education meta-universe creates digital identities for social education-related subjects (without age, class, and cultural level restrictions), and provides a social education environment that integrates virtuality and reality. With citizens as the center and digital twin technology as the support, metaverse provides hardware resources, educational resources, human resources and so on. It integrates technology, resources, data and services. It is perceptive, adaptive, interactive and ubiquitous.

Metaverse enabling intelligence education needs a lot of data transmission, storage, and technical support, including the low delay and high-rate characteristics of 5G, digital twin, artificial intelligence and blockchain [30,36,41,50]. The metaverse endows an open environment for intelligent education, transcends the limitations of time and space, realizes free and open innovation and creation at the same time, and transforms the cultivation mode of innovative education and innovative thinking. The metaverse empowers the sense of presence of intelligent education and allows the educated subject of society to simulate the three-dimensional interaction of the real world with a three-dimensional image. The intelligent education environment empowered by the meta-universe based on intelligent technology can meet the needs of personalized learning of citizens. It allows citizens to realize autonomous learning and personalized learning according to their own learning and cognitive style, measure and diagnose the existing level of citizens, and collect and analyze the learning data of citizens. Metaverse Enabling Library social education also shows its potential to plan learning content based on the analysis of citizens' individual needs and learning level, to provide citizens with immersive interactive experience and make the virtual reality. The metaverse uses XR (VR/AR/MR) technology to immerse learners in

the learning content, and can also guide citizens to actively develop related functions in the metaverse to improve the learning effect and improve the game-based language learning mode.

Self-powered sensors have broad application prospects because of their flexibility and no external power supply. At present, research at home and abroad is mostly concentrated in high-end fields such as wearable devices and electronic skin. In addition, it also has good research potential in the fields of environmental monitoring, sports, and games. In recent years, machine learning, especially deep learning, has gradually become a research hotspot and mainstream development direction in the field of Internet of Things and sensors. The combination of big data and machine learning technology with self-powered sensor data processing provides a new idea for researchers. As the resources and tools of machine learning are abundant and easy to obtain, it has more efficient data processing capabilities, which can make sensors intelligent, and save testing time and costs. Compared with traditional sensors, the data generated by self-powered sensors are more suitable for processing and analysis using machine learning techniques, so machine learning has gradually appeared in the research field of self-powered sensors.

The rise of the metaverse brings new opportunities for smart education. In this paper, we combine deep learning, self-powered sensors, and data analysis technology to study a new teaching management mechanism to achieve classroom attendance through intelligent methods. Neural networks are a type of machine learning model that is inspired by the structure and functioning of the human brain. They consist of interconnected nodes, called neurons, organized in layers. Each neuron performs a simple computation and passes its output to other neurons. By adjusting the connections between neurons, neural networks can learn to recognize patterns, make predictions, or perform other tasks. Neural networks have gained popularity because of their ability to learn and generalize from large amounts of data. Specifically, we first take intelligent self-powered sensors to collect data from remote scenarios and send them back. Then, we employ the deep learning model to design a face recognition algorithm to automatically detect and identify students and teachers, and finally realize the intelligent attendance task in classroom education.

The main contributions of our work are summarized as follows.

(1) We propose a framework for teaching management mechanisms combining deep learning and self-powered sensors. This mechanism explores the new opportunities and development ideas brought to education management under the background of metaverse space.

(2) We propose a multi-level features fusion face recognition model that combines traditional methods and deep learning methods. We employ sparse coding to extract primary features of faces and utilize a convolutional neural network model to extract and learn high-level feature knowledge vectors in images, which greatly improves the recognition and classification performance of the model.

The overall structure of the article is arranged as follows. Section 2 describes the related work about some current methods of face recognition and metaverse. Section 3 details the face recognition method for smart teaching management mechanism method proposed in this paper. Section 4 verifies the effectiveness of the method from the experimental point of view. Section 5 concludes the whole paper.

2. Related work

In this section, we introduce some current methods of face recognition, including traditional methods and deep learning methods. Face recognition is an important intelligent target recognition technology, which can use the visual feature information of face images for identity recognition, and has a very broad development prospect [21, 48]. Face recognition is one of the most natural and visual recognition methods and can be widely used in many important industries, such as public security, enterprises, airports, supermarkets, military and aerospace.

The method based on geometric features appeared in the early stage of face recognition research. It analyzes the shape and geometric relationship of facial organs such as eyes, nose, mouth, and so on, and then identifies them. Face recognition based on geometric features is intuitive and fast, but its robustness to illumination, expression, pose, and occlusion is poor. Moreover, the recognition rate of the algorithm is relatively low because it only uses the geometric features of the image and ignores the local details. At present, such methods are rarely used on their own, and usually only as a useful supplement to other identification methods. Subspace-based method is the most common face recognition method. It transforms the original data into a low-dimensional subspace and makes the distribution of data in this subspace more compact, which not only reduces the computational complexity, but also makes the classification more accurate. The space transformation includes linear transformation and nonlinear transformation. Principal Component Analysis (PCA) is the most widely used feature extraction method. It finds a set of optimal unit orthogonal vector basis through linear transformation, and uses their linear combination to reconstruct the image, to minimize the error between the reconstructed image and the original image [44]. This method can describe samples well with fewer features, but it does not consider the differences between samples of different categories, and the features are not sufficient for classification. Linear Discriminant Analysis (LDA) takes the separability of samples as the goal, and seeks a set of linear transformations so that similar samples can be concentrated and dissimilar samples can be separated as much as possible after dimensionality reduction [2]. This method has a good recognition effect when the sample is separable, but the recognition rate is low when the data is not linearly separable. The nonlinear analysis method includes two kinds: the kernel-based method and the manifold learning method. At present, nonlinear analysis methods such as kernel PCA, kernel Fisher discriminant analysis and kernel ICA have been successfully applied to face recognition. However, it is very difficult to select kernel function and its parameters in practical application, and this kind of method is not suitable for large training samples because the vector dimension will become higher after kernel mapping [20,27]. The nonlinear analysis method based on manifold can find the embedded low-dimensional manifold by observing the high-dimensional input space and recover the intrinsic structure of the data in the low-dimensional space. At present, isometric feature mapping method, local linear embedding method and the Laplacian feature mapping method are the three manifold learning methods that are most widely used and have the greatest influence. However, these methods have a heavy computational load in practical application, and due to the lack of clear projection matrix, they cannot establish mapping relations for new test samples [3, 35, 43].

Face recognition based on learning is highly targeted to specific applications and plays an important role in the field of face recognition. Inspired by the sparse coding in human visual cortex, in [31], the authors proposed a more effective representation method for signals in 1996, namely sparse representation, which could provide a good description of signals by the linear combination of a few dictionary atoms, that is to say, a few resources were used to represent important information of interest. In 2009, the authors of [46] applied sparse representation theory to face recognition technology for the first time. They represented the face to be tested by a linear combination of the training sample set, and then judged the identity of the test sample according to the reconstruction error with each category. Different from the traditional face recognition methods, this method is not sensitive to feature selection. When a certain feature dimension is reached, it can achieve good recognition effect even with down-sampling and has good robustness to illumination, expression, camouflage and small pose changes.

At present, the development of traditional face recognition algorithms has basically become mature and perfect, and the classical algorithms such as Eigenfaces and Fisherfaces [22] have been born and become the Benchmark algorithms for commercial applications. The study of face recognition technology in natural scenes is a new challenge in the field of face recognition.

Because the natural environment conditions are closer to the real application scenarios, the study of face recognition technology in natural scenes has more realistic significance in application.

With the gradual evolution of deep learning methods, the probability that the facial features extracted by deep learning are affected by objective factors has dropped sharply. The facial feature extraction method based on deep learning has achieved 98% accuracy on the LFW dataset. Image-based face recognition technology has matured in the experimental environment. However, in the actual classroom environment video streaming, there are still some problems.

In [11], the authors proposed a large-scale feature search method, which can generate new and more powerful feature expressions composed of many complex structures, nonlinear properties, and multi-layer neural morphological structures. In this method, features are randomly generated and filtered out for optimal expression to fit a specific task. In [18], the authors proposed the usage of deep convolutional confidence network to learn and obtain hierarchical expression. By using the global structure of an object class, a new convolution learning model is proposed, which simulates the global structure without assuming the stability of features on the whole image while maintaining the robustness of the algorithm under slight registration imbalance. In [40], the authors proposed a hybrid convolutional networks-restricted Boltzmann machine model for face recognition in unrestricted environments. A key point of this method is to directly learn interrelated visual features that indicate identity similarity from the pixels of the original face image pair through a hybrid deep network. In this model, deep convolutional network simulates human primary visual cortex to extract locally associated visual features from two face images, and then further extracts high-order global features through multiple network layers. There is an RBM model at the top level, which further processes the complementary features learned by different convolutional network groups through a two-level average pooling process. The resulting model has strong performance. In [39], the authors proposed a face recognition algorithm based on deep learning and joint features within and between classes, which put face images into convolutional neural networks. They used the designed cost function to train the network, and used the identification and verification of two supervision signals to increase the difference between classes and reduce the differences within the class, so that the obtained features are more conducive to identification. The experimental results show that the detection rate of the algorithm on the LFW face database has reached an unprecedented level of recognition.

There are also many applications of face recognition in classroom attendance. In [47], the authors used face recognition technology to realize the function of timed attendance of students in class. The system uses face information as the basis for attendance. Compared with fingerprint recognition and iris recognition, it is easy to obtain, intuitive and friendly, and easy to distinguish. However, this system has problems such as low multi-face detection accuracy and high face recognition misjudgment rate, and the extracted facial features are easily affected by illumination, posture, and expression. In [6], the authors proposed proposes a classroom face recognition system based on Android mobile terminals, which uses haar face detection method and VGG face feature extraction network method to identify students' faces collected by mobile phone cameras. However, due to the limited shooting area of the system, it did not play a role in urging students to attend classes. In [37], the authors uses the AdaBoost face detection algorithm to obtain images of a single face area, and implements the classroom attendance system through the principal component analysis algorithm (Principal Component Analysis, PCA), but the PCA algorithm is more sensitive to conditions such as light, age, and expression. The consistency of the extracted facial feature information cannot be guaranteed and the recognition effect is not good.

While the concept of the metaverse is still evolving, there is growing interest in exploring its potential applications in various fields [13, 45], including educational administration. One relevant work in the literature is the research in [9], that investigate the integration of the metaverse in educational administration to enhance collaboration and engagement among students and administrators. They propose a framework that leverages virtual reality (VR) technologies

within the metaverse to create immersive educational environments, enabling administrators to remotely manage educational processes, facilitate communication, and provide personalized support to students. Another study examined the potential of the metaverse in reimagining educational leadership [5]. In that paper, the authors discussed how immersive experiences and virtual simulations in the metaverse can be used to develop leadership skills, promote innovative decision-making and foster collaboration among educational leaders. These studies highlight the emerging exploration of the metaverse's application in educational administration, paving the way for future research and implementation in this domain.

3. Proposed method

This section describes in detail the face recognition method for smart teaching management mechanism proposed in this paper.

Based on the shortcomings of the above methods, we combine traditional methods and deep learning methods to propose a multi-level features fusion face recognition model. After collecting data through self-powered sensors, we transmit the collected data back to the server. We first employ sparse coding to extract primary features of faces. Second, we utilize a convolutional neural network model to extract and learn high-level feature knowledge vectors in images. Finally, the face is recognized and classified based on the extracted feature knowledge vector.

3.1. **Data collection with self-powered sensors.** In our approach, we first utilize self-powered sensors to collect student data in playground, outdoor sports and remote scenarios. Currently, self-powered sensors include sound sensors, pressure sensors, moving object self-driven sensors, vibration sensors, self-driven environmental sensors and biomedical sensors. We utilize self-powered sensors to collect image information, spatial location information, temperature information, and frequency information. The collected data comes in various forms. With the rapid development of machine learning technology, more and more people use machine learning technology to process massive and complex sensor data. The machine learning workflow can be simply described as data preprocessing, feature extraction, training model, validation. In our approach, our overall framework for collecting and processing data by combining self-powered sensors and deep learning models is shown in Fig.1. First, self-powered sensors collect data and transmit it to a data center for storage and management. Then, we send the selected sensor data to the server through the control center. Next, we perform low-level and high-level feature extraction and fusion on the data to train a machine learning model. Finally, the results are validated and evaluated by the machine learning model.

After the data collection is completed, next, we will introduce the feature extraction process for the collected data, including low-level feature extraction, high-level feature extraction, and model training.

3.2. **Low-level feature extraction.** In recent years, the usage of sparse coding theory to conduct face recognition research is a popular direction. The sparse coding is an essential aspect of sparse representation theory. Researchers have done a lot of research work on how to obtain the sparsest coefficient solution under an over-complete dictionary. Later, scholars conducted in-depth research on the sparse optimization problem and proposed many effective sparse coding algorithms, which are mainly divided into two categories: greedy algorithms and relaxed optimization algorithms. The greedy algorithm finds the dictionary atom that best matches the signal in an iterative manner, thereby constructing a sparse approximation of the signal. The MP algorithm is a pure greedy algorithm, which has the advantages of simple structure and high efficiency, but because the projection of the signal residual on the selected atom set is not an orthogonal projection, the dictionary atoms obtained by each iteration are not necessarily optimal, and the convergence is very slow [25]. On this basis, the OMP algorithm adds orthogonalization operations in the iterative process, which improves the accuracy of sparse

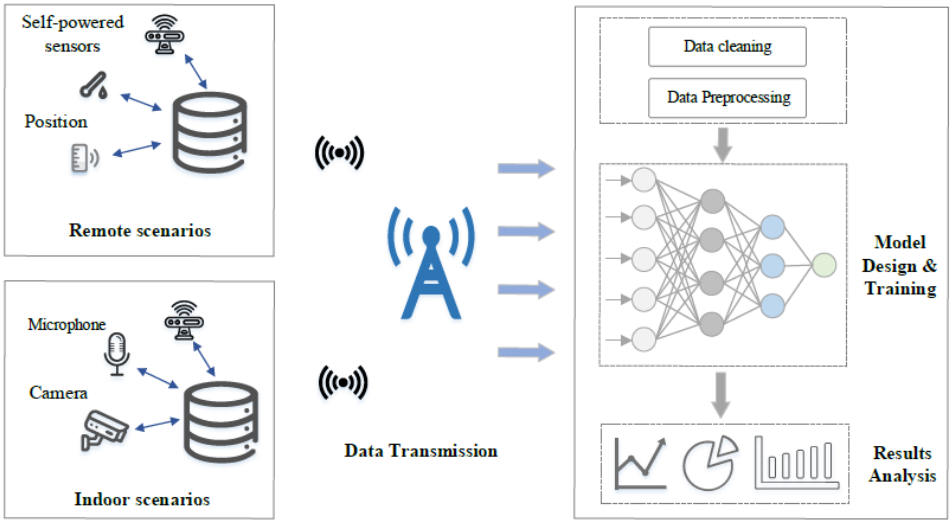


Figure 1. The overall framework of the proposed method for collecting and processing data by combining self-powered sensors and deep learning models.

representation and the speed of convergence [32]. The relaxation optimization algorithm relaxes the non-convex L_0 norm to a convex L_1 norm or other more tractable sparsity and then solves the combinatorial optimization problem through convex optimization or nonlinear optimization. The commonly used relaxation optimization algorithms include Basis Pursuit (BP) and Basis Pursuit Denoising (BPDN). The BP algorithm and the BPDN algorithm use the L_1 norm instead of the L_0 norm, solving it by the linear programming. Compared with the greedy algorithm, the sparse representation obtained by the relaxation optimization algorithm is more accurate, but the complexity is relatively high [7], so the greedy algorithm has strong practicability in some cases where the accuracy requirement is not high.

In general, most feature learning methods try to model the real distribution of the given training data. In other words, feature learning is a learning model that describes an approximation of the true distribution of the data. These methods include denoising autoencoders, restricted Boltzmann machines (RBMs), independent component analysis (ICA) and sparse coding. These methods work well, but all of them have a disadvantage. All of them need to adjust a lot of parameters, such as learning rates, momentum penalties, sparsity penalties and weight decay. The final determination of these parameters needs to be obtained by cross validation, which takes a long time to train such a structure, so it takes more time to obtain so many parameters by cross validation. It takes a lot of effort to adjust to get a good set of parameters. Although ICA only needs to adjust one parameter, it has weak expansion ability for high-dimensional input or large feature sets. The sparse filtering that is a simple feature sparse optimization method, is proposed to solve the problem of cumbersome parameter regulation in sparse coding. Although it has only one hyperparameter to tune, it works well. Different from other feature learning methods, sparse filtering does not explicitly build the distribution model of input data and avoids explicit modeling of data distribution. Instead, it optimizes the sparsity of feature distribution to obtain good feature expression. Sparse coding needs to find active basis coefficients in each iteration, which is time-consuming. Moreover, the sparsity factor is also a parameter that needs to be adjusted.

To learn features in images, we first introduce a feature distribution matrix M, where each row of the matrix is a feature and each column is a sample. Each element $h_j^{(i)}$ represents the activation value of the j-th feature of the i-th sample. Next, consider computing linear features from each sample. Specifically, denote the j-th eigenvalue (j-th row in the feature matrix) of the i-th sample (i-th column in the feature matrix) by $h_j^{(i)}$. Because it is a linear feature, it can

be calculated by the following formula

$$h_j^{(i)} = w_j^T x^{(i)}.$$

In the first step, we simply normalize the rows of the feature matrix, then the columns and finally sum the absolute values of all elements in the matrix. Specifically, each feature is first normalized to equal activation values. This is done by dividing each feature by its L_2-norm across all samples

$$\overrightarrow{h_j} = \frac{h_j}{\|h_j\|_2}.$$

Then, normalize the features of each sample. Thus, they will fall on the 2-norm unit sphere unit L_2-ball. They can be calculated as follows

$$\widehat{h} = \frac{\overrightarrow{h_j}}{\left\|\overrightarrow{h^{(l)}}\right\|_2}.$$

These normalized features are then optimized using an L_1 norm penalty to constrain sparsity. For a dataset with M samples, the objective function of sparse filtering is expressed as:

$$\text{Minimize} \sum_{i=1}^{M} \left\|\widehat{h}^{(i)}\right\|_1 = \sum_{i=1}^{M} \left\| \frac{\overrightarrow{h_j}}{\left\|\overrightarrow{h^{(i)}}\right\|_2} \right\|.$$

This item measures the population sparsity of the features of the $i-th$ sample, that is, limiting each sample to only a few non-zero values. Because the normalized feature $\hat{h}^(i))$ is constrained to only fall on the unit sphere of L_2 norm. So, when these features are sparse, that is, when the sample is close to the feature coordinate axis, the above objective function will be the smallest. On the contrary, if each eigenvalue of a sample is similar, it will lead to a high penalty. In sparse filtering, we introduce competition to the filters, replacing the general L_1 norm penalty with a normalized sparsity penalty (like L_1/L_2), which can replace the mutual orthogonality constraint between filters in ICA. After the algorithm obtains the objective function, the L-BFGS function optimization package can be easily used to optimize the solution, and the soft-absolute function is usually used as the feature activation function in the experiment as follows.

$$h_j^{(i)} = \sqrt{\varepsilon + \left(W_j^T X^{(i)}\right)^2} \approx |W_j^T X^{(i)}|.$$

3.3. **High-level feature extraction.** Deep learning is a class of representation learning methods in machine learning. An observation can be represented in several ways, as a vector of intensity values for each pixel, or more abstractly as a series of edges, a region of a particular shape [38]. Distributed representation assumes that observations are generated by the interaction of different factors. On this basis, deep learning further assumes that the process of this interaction can be divided into multiple levels, representing multiple layers of abstraction over observations. Different numbers of layers and scales of layers can be used for different levels of abstraction. Deep learning uses this idea of hierarchical abstraction, where higher-level concepts are learned from lower-level concepts. This hierarchical structure is often constructed layer by layer using a greedy algorithm to select more efficient features for machine learning. Many deep learning algorithms come in the form of unsupervised learning, so these algorithms can be applied to unlabeled data that other algorithm cannot. This type of data is richer and easier to obtain than labeled data. This also gives an important advantage to deep learning [4].

We design a high-level feature extraction model based on convolutional neural networks (CNN). Combined with the low-level features extracted by sparse coding, we fuse multi-level

feature knowledge vectors for face recognition. The model structure of the proposed method is shown in Fig.2.

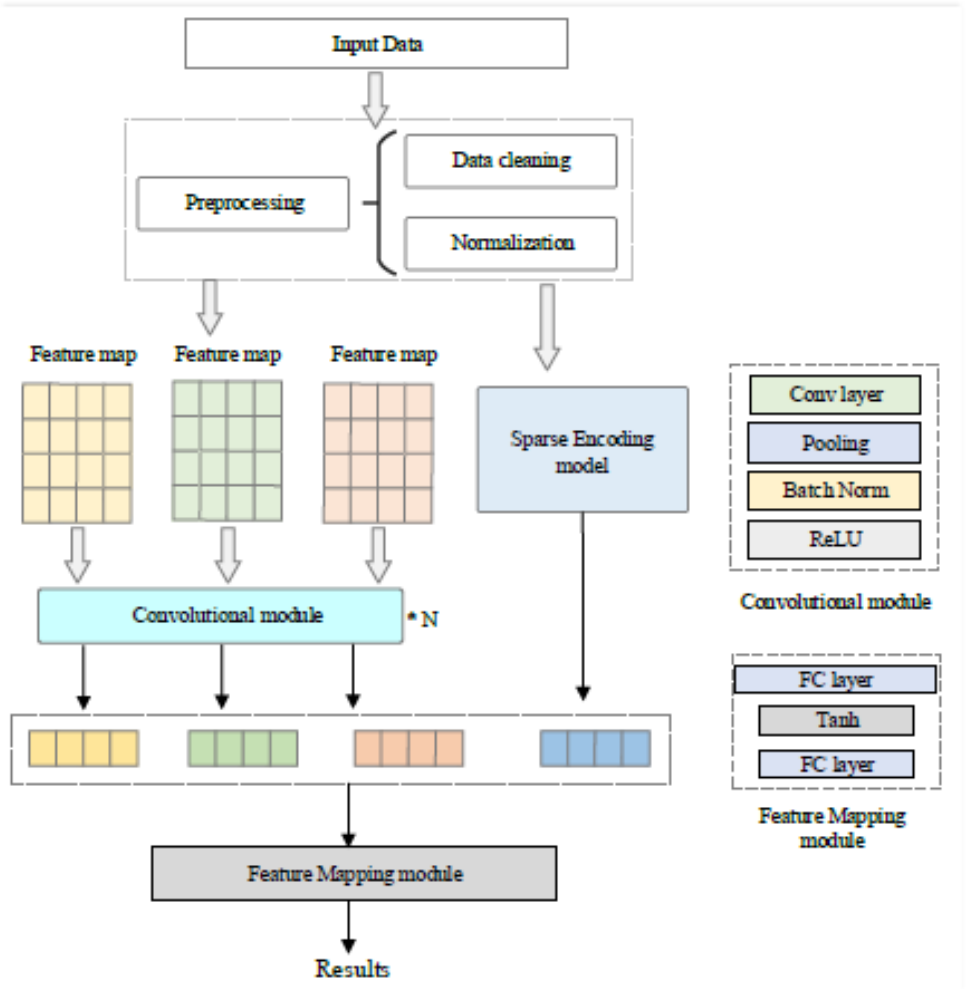


Figure 2. The model structure of the proposed method.

The CNN is partially similar to early hierarchical visual models based on human visual system, such as the alternating structure of S-layer and C-layer. However, CNN does not rely on manual selection of features, but constantly learning features. Convolutional neural networks achieve invariance of displacement, scaling, and distortion by combining three methods: Local receptive field, weight sharing and sub-sampling. Local receptive field refers to that each neuron is only connected with neural units in a small neighborhood of the previous layer, and some features in the neighborhood can be extracted. Weight sharing means that the connection weights of neurons on the same feature map and the local receptive field of the previous layer are the same, and the weights of different feature maps are different. Weight sharing makes the convolutional neural network have fewer parameters. Neurons in the secondary sampling layer sample the data in their receptive areas, so that the number of neurons in the feature plane of the secondary sampling layer is halved.

The structure of convolutional neural network mainly includes three layers: Feature extraction layer, pooling layer, and feature mapping layer.

(1) Feature extraction layer. Each neuron obtains the input value from the convolution window of the previous layer, and extracts the local feature by the convolution operation between its weight and the input. The weight is constantly updated in the learning process, rather than specified in advance.

(2) sub-sampling layer. There is a sub-sampling layer behind each convolutional layer, which is used to reduce the resolution of the feature map and the influence of translation and other deformation on the feature map.

(3) Feature mapping layer. Each computational layer of convolutional network is composed of multiple feature maps, and each feature map is a plane with equal neuron weights, which can reduce the number of free parameters of the network. The feature map structure adopts nonlinear activation function, such as Sigmoid function, so that the feature map has translation invariance. Each neuron of the convolutional layer acts as a filter to extract the local features of the feature map of the previous layer. The first layer of convolution is to extract the features of the input image, and the feature map is obtained after convolution, which is used as the input of the next layer, that is, the sub-sampling layer. Like the convolutional layer, the neurons of each feature map in the pooling layer also share the connection weight, and each neuron receives data from its receptive domain. Each feature map of the convolution layer corresponds to a feature map of the sub-sampling layer. The last layer is the output layer, the number of nodes in the output layer represents the number of classifications, and the previous layer of the output layer is the fully connected layer.

3.4. **Model training.** After the model design is completed, we need to train the network model to obtain the optimal parameters. The training method of convolutional neural network is generally similar to that of traditional neural network, but due to its special structure, different layers have different error return and forward activation. The specific steps are as follows. We select a sample (x_i, y_i) from the sample set, input the x_i vector into the network, and calculate the corresponding actual output $f(x_i)$. We define the loss function as follows.

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i log f(x_i) + \lambda \sum_{j=1}^{n} ||W_j||^2,$$

where m is the number of samples in the sample set, and j is the number of layers of the convolutional neural network. The network parameter update needs to calculate the network parameter residual, which includes the residual of the output layer, the residual of the convolutional layer whose next layer is the pooling layer, and the residual of the pooling layer whose next layer is the convolutional layer. Then, take the partial derivative with respect to the parameter W, as follows.

$$\frac{\partial L}{\partial W} = -\frac{1}{m}\big(a(y) - f(x)\big) * f(x)' + \lambda W_l,$$

where $a(y)$ represents the one-hot representation of the label for the sample x, that is, only one element is 1, and the rest are 0.

4. Experiments and results

In this section, we evaluate the proposed a multi-level features fusion face recognition model based on the traditional methods and deep learning model in this paper. First, we describe the detailed characteristics of the datasets used in the experiment and the evaluation indicators used in the model validation. Secondly, we compare and analyze the proposed method with the current face recognition algorithms. Third, we conduct ablation experiments on sparse coding methods and deep learning methods to verify the performance of each module designed in this paper.

4.1. **Indicators and dataset.** The performance indicators adopted in our experiments are the common metrics in face recognition research, including Recall, Precision, Accuracy, and F1-score. They are respectively defined as follows.

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$
$$\text{Accuracy} = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$
$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}.$$

Among them, TP means the true positive prediction, TN means the true negative prediction, FP means the false positive prediction, FN means the false negative prediction. Each indicator represents the measurement of the error between the predicted label and the actual label of data label. The more accurate the predicted value of evaluation is, the lager each indicator is, and the better the performance of the model is.

In the experiment, we use two face databases to verify the model performance, including FERET [33] and LFW-a databases [49]. The FERET face database contains 1400 images of 200 individuals, each of whom has 7 face photos in different poses. The multiple face photos of each person in the database contain face pose, expression, and lighting variation factors. The LFW-a face database consists of 13,233 images belonging to 5,749 individuals. These face pictures contain many factors that interfere with recognition, including complex backgrounds, occlusions, expressions, and large changes in posture. Therefore, it has become the preferred database for studying face recognition algorithms in natural scenes. In addition, LFW-a is a variant of the LFW database, the only difference from LFW is that the images are aligned with commercial software, and the color images are converted to grayscale images. Using the LFW-a database can improve the algorithm performance and have a higher recognition rate than using the original LFW database. The original LFW database is divided into two parts, view1 and view2. view1 is used for model selection and algorithm development, and view2 is used to evaluate algorithm performance and submit algorithm performance reports to the database management group.

4.2. **Performance comparison.** We compare and analyze the method proposed in this paper with the current face recognition algorithms on the above two datasets. We compare some classic machine learning algorithms and deep learning face recognition models, including support vector machines (SVM), decision trees (DT), fully connected networks (FC network), Spartans [19], Hierarchical-PEP [24], and the YoLo model [8]. Table 1 shows the comparison results of our proposed method with other methods on the FERET dataset. Table 2 shows the comparison results of our proposed method with other methods on the LFW-a dataset.

Table 1. The comparison results of our proposed method with other methods on the FERET dataset.

Methods	Recall (%)	Precision (%)	Accuracy (%)	F1-score (%)
DT	71.51	73.19	78.64	72.34
SVM	78.32	79.42	81.37	78.87
FC network	85.67	82.87	89.64	84.25
Spartans	86.36	89.54	91.36	87.92
Hierarchical-PEP	89.24	86.34	92.34	87.77
YoLo	90.44	90.06	93.76	90.25
Ours	89.68	91.38	94.35	90.52

It can be seen from the results in the two tables: (1) Compared with the machine learning model, the performance of the algorithm based on the neural network model exceeds the performance of the machine learning model. For example, the accuracy of SVM and DT is less than 82% on the FERET dataset. This shows that the neural network model has better feature extraction ability than the traditional method. (2) In the neural network model, our method achieves optimal results, such as 90.52% in the F1-score on the FERET dataset, which proves that our proposed model has good performance on face recognition tasks. In addition, our face recognition model achieves more than 93% accuracy on both datasets, which also implies that the proposed model has a good performance.

Table 2. The comparison results of our proposed method with other methods on the LFW-a dataset.

Methods	Recall (%)	Precision (%)	Accuracy (%)	F1-score (%)
DT	67.31	65.12	70.24	66.2
SVM	73.64	72.46	73.31	73.05
FC network	85.18	83.64	87.46	84.4
Spartans	88.19	84.76	89.24	86.44
Hierarchical-PEP	90.16	88.46	91.03	89.3
YoLo	91.08	91.27	92.78	91.17
Ours	92.35	91.15	93.41	91.75

4.3. **Ablation study.** In this section, we conduct ablation experiments on sparse coding methods and deep learning methods to verify the performance of each module designed in this paper. First, we change the hidden layers in the sparse coding and deep learning models to verify the impact of the number of hidden layers on accuracy. Second, we verified the impact of multi-layer feature fusion on model performance, that is, we ablated low-level features and high-level features to see the changes in model performance. Third, we verified different activation functions in the deep learning model, including the sigmoid function, the Tanh function, the ReLU function, and the Softmax function.

Table 3. The ablation results for multi-level features on the LFW-a dataset.

Methods	Recall (%)	Precision (%)	Accuracy (%)	F1-score (%)
Low-level features	70.62	71.85	72.52	71.23
High-level features	87.25	86.54	89.42	86.89
All	92.35	91.15	93.41	91.75

Table 3 demonstrates the role of multi-level features. It can be seen from the results in the table that only using low-level features makes the performance of the model drop a lot, for example, the accuracy rate drops by 20.89%. When using high-level features, the performance of the model does not drop much, which shows that the deep learning model has a strong ability to extract data features. In addition, we can also see that the combination of low-level features and high-level features improves the performance of the model, which also reflects the effectiveness of the multi-level feature fusion model proposed in this paper in data feature extraction.

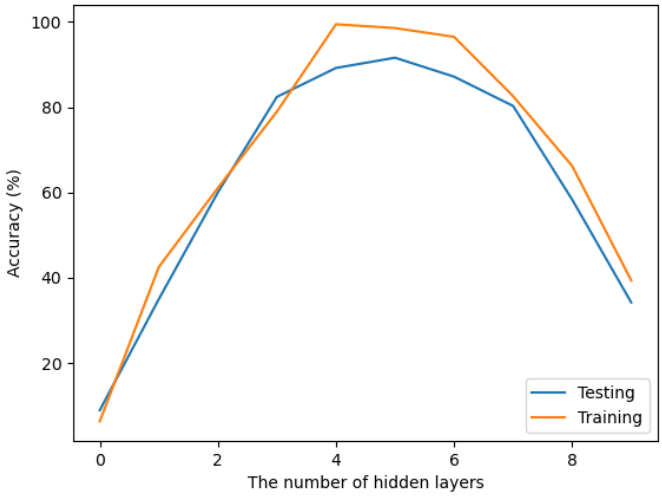


Figure 3. The accuracy of the proposed model on the FERET dataset with the change of the number of hidden layers.

Fig.3 shows the variation of accuracy of the proposed model with the change of the number of hidden layers. We can see two curves: training and testing. The training and testing curves

firstly increase and then decrease with the increase of the number of hidden layers. The accuracy of the highest peak reached more than 90%. This experimental phenomenon shows that it is not that the more hidden layers, the better, but that the model can show the best performance when the hidden layers are 4 to 6. And, we can also see that the training curve is very close to the test results at the beginning.

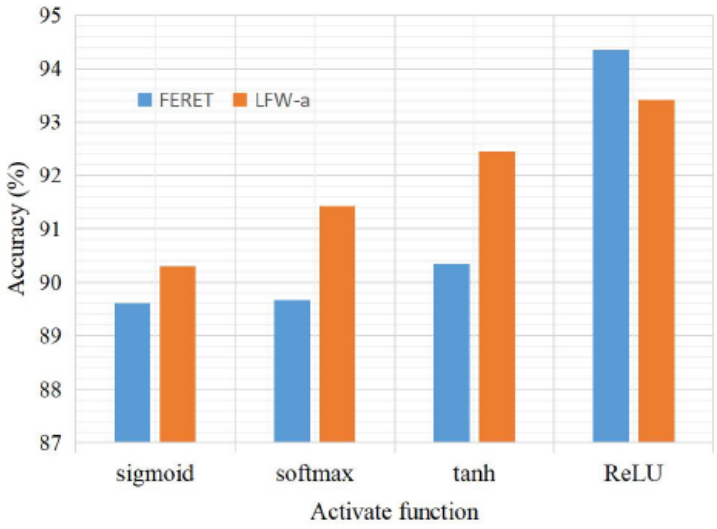


Figure 4. The comparison of different activate functions on two datasets.

Fig.4 shows the accuracy results of different activation functions on the two datasets. The blue bars represent the model's accuracy results on dataset FERET, while the yellow bars represent the model's accuracy results on dataset LFW-a. Among the four activation functions, the ReLU function shows the best performance on both datasets, e.g., the accuracy is 94.35% on the FERET dataset. However, using the sigmoid function to calculate the error gradient in backpropagation requires a large amount of computation, while using the ReLU activation function saves a lot of computation in the entire process. For a deep network, when the sigmoid function is back-propagated, it is easy for the gradient to disappear, so that the training of the deep network cannot be completed. ReLU will make the output of some neurons to be 0, which causes the sparsity of the network, reduces the interdependence of parameters, and alleviates the overfitting problem. This is also the reason why the ReLU function is widely used.

5. Conclusion

This paper combines the self-powered sensor and deep learning model to design a face recognition algorithm to automatically detect and identify students and teachers, and finally realize the intelligent education with metaverse. This article is novel in two aspects: one is that we combine the metaverse to make education more intelligent, and the other is that we take advantage of the self-powered sensors to achieve data collection intelligently. We analyze the importance of face recognition technology in the application of intelligent education scenarios. We combine traditional learning methods and deep learning methods to propose a deep learning face recognition network model with improved multi-level feature fusion, and comprehensively describe its key network structure design and algorithm. First, we first take intelligent self-powered sensors to collect data from remote scenarios and send them back. Second, we employ the sparse coding algorithm to extract primary low-level features of face images. Third, we utilize a convolutional neural network model to extract and learn high-level feature knowledge vectors in face images. Finally, the face is recognized and classified based on the extracted feature knowledge vector. We evaluate our proposed method on two public datasets, i.e., the FERET face database and

the LFW-a database. We compare the method with the current face recognition algorithms on the above two datasets and the results show that the performance of our method is superior to other face recognition method. In the experiments, we conduct ablation experiments to verify the effect of the number of hidden layers and different activation functions on the performance of the model. The results of the ablation study show that when the number of hidden layers is 5 and the activation function is the ReLU, the performance of the model is optimal. Future work will focus on applying metaverse and deep learning technologies to more educational scenarios and exploring more data feature extraction and fusion models to improve method performance.

6. Acknowledgment

This paper was supported by Jilin Province Science and Technology Development Plan Project (Granted No.s 20230203035SF, YDZJ202303CGZH001, YDZJ202201ZYTS605, and 20200404223YY), Research Topic Project of Education and Teaching in Beihua University (Granted No. 20230006) and National Natural Science Foundation of China (Granted No. 62202247).

References

[1] Allam, Z., Dhunny, Z.A. On big data, artificial intelligence and smart cities, *Cities*, V.89, 2019, pp.80-91.
[2] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, V.19, N.7, 1997, pp.711-720.
[3] Belkin, M., Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, V.15, N.6, 2003, pp.1373-1396.
[4] Bengio, Y., Courville, A., Vincent, P. Representation learning: A review and new perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, V.35, N.8, 2013, pp.1798-1828.
[5] Breitkopf, D., Albold, C., Barlow, J., et al. Virtual residency recruitment: future directions in the new era, *Annals of Medicine*, V.54, N.1, 2022, pp.3341-3347.
[6] Chen, H., Haoyu, C. Face recognition algorithm based on VGG network model and SVM, *Journal of Physics: Conference Series, IOP Publishing*, V.1229, N.1, 2019, pp.012015.
[7] Chen, S.S., Donoho, D.L., Saunders, M.A. Atomic decomposition by basis pursuit, *SIAMReview*, V.43, N.1, 2001, pp.129-159.
[8] Chen, W., Huang, H., Peng, S., et al. YOLO-face: a real-time face detector, *The Visual Computer*, V.37, N.1, 2021, pp.805-813.
[9] Chen, Y., Lin, W., Zheng, Y., et al. Application of active learning strategies in metaverse to improve student engagement: An immersive blended pedagogy bridging patient care and scientific inquiry in pandemic, *Available at SSRN 4098179*, 2022, pp.1-20.
[10] Çlötekin, A., Lochhead, I., Madden, M., et al. Extended reality in spatial sciences: A review of research challenges and future directions, *ISPRS International Journal of Geo-Information*, V.9, N.7, 2020, pp.439.
[11] Cox, D., Pinto, N. Beyond simple features: A large-scale feature search approach to unconstrained face recognition, *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE, 2011, pp.8-15.
[12] Da, X.L., He, W., Li, S. Internet of Things in industries: A survey, *IEEE Transactions on Industrial Informatics*, V.10, N.4, 2014, pp.2233-2243.
[13] Deveci, M., Gokasar, I., Castillo, O., et al. Evaluation of metaverse integration of freight fluidity measurement alternatives using fuzzy Dombi EDAS model, *Computers & Industrial Engineering*, V.174, 2022, pp.108773.
[14] Díaz, J., Saldaña, C., Avila, C. Virtual world as a resource for hybrid education, *International Journal of Emerging Technologies in Learning*, V.15, N.15, 2020, pp.94-109.
[15] Dionisio, J.D.N., Burns III W.G., Gilbert, R. 3D virtual worlds and the metaverse: Current status and future possibilities, *ACM Computing Surveys (CSUR)*, V.45, N.3, 2013, pp.1-38.
[16] Farshid, M., Paschen, J., Eriksson, T., et al. Go boldly!: Explore augmented reality (AR), virtual reality (VR), and mixed reality (MR) for business, *Business Horizons*, V.61, N.5, 2018, pp.657-663.
[17] Gadea, C., Hong, D., Ionescu, D., et al. An architecture for web-based collaborative 3D virtual spaces using DOM synchronization, *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, IEEE, 2016, pp.1-6.

[18] Huang, G.B., Lee, H., Learned-Miller, E. Learning hierarchical representations for face verification with convolutional deep belief networks, *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp.2518-2525.
[19] Juefei-Xu, F., Luu, K., Savvides, M. Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios, *IEEE Transactions on Image Processing*, V.24, N.12, 2015, pp.4780-4795.
[20] Kim, K.I., Jung, K., Kim, H.J., Face recognition using kernel principal component analysis, *IEEE Signal Processing Letters*, V.9, N.2, 2002, pp.40-42.
[21] Kortli, Y., Jridi, M., Al Falou, A., et al. Face recognition systems: A survey, *Sensors*, V.20, N.2, 2020, pp.342.
[22] Kshirsagar, V. P., M. R. Baviskar, and M. E. Gaikwad. Face recognition using Eigenfaces, *3rd International Conference on Computer Research and Development, IEEE,* V.2, 2011, pp.302-306.
[23] Lasi, H., Fettke, P., Kemper, H.G., et al. Industry 4.0, *Business & Information Systems Engineering*, V.6, N.4, 2014, pp.239-242.
[24] Li, H., Hua, G. Hierarchical-pep model for real-world face recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp.4055-4064.
[25] Mallat, S.G., Zhang, Z. Matching pursuits with time-frequency dictionaries, *IEEE Transactions on Signal Processing*, V.41, N.12, 1993, pp.3397-3415.
[26] McNeil, S.E. Nanotechnology for the biologist, *Journal of Leukocyte Biology*, V.78, N.3, 2005, pp.585-594.
[27] Mika, S., Rätsch, G., Müller, K.R., A mathematical programming approach to the Kernel Fisher Algorithm, *Advances in Neural Information Processing Systems*, 2000, pp.13.
[28] Milgram, P., Takemura, H., Utsumi, A., et al. Augmented reality: A class of displays on the reality-virtuality continuum, *Telemanipulator and Telepresence Technologies*, SPIE, V.2351, 1995, pp.282-292.
[29] Mystakidis, S., Metaverse, *Encyclopedia*, V.2, N.1, 2022, pp.486-497.
[30] Ng, W.C., Lim, W.Y.B, Ng, J.S., et al. Unified resource allocation framework for the edge intelligence-enabled metaverse, *IEEE International Conference on Communications*, 2022, pp.5214-5219.
[31] Olshausen, B.A., Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, V.381, N.6583, 1996, pp.607-609.
[32] Pati, Y.C., Rezaiifar, R., Krishnaprasad P S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, IEEE, 1993, pp.40-44.
[33] Phillips, P.J., Wechsler H., Huang J., et al. The FERET database and evaluation procedure for face-recognition algorithms, *Image and Vision Computing*, V.16, N.5, 1998, pp.295-306.
[34] Qian, L., Luo, Z., Du, Y., et al. Cloud computing: An overview, *IEEE International Conference on Cloud Computing*, Springer, Berlin, Heidelberg, V.5931, 2009, pp.626-631.
[35] Roweis, S.T., Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding, *Science*, V.290, N.5500, 2000, pp.2323-2326.
[36] Shafique, K., Khawaja, B.A., Sabir, F., et al. Internet of Things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios, *IEEE Access*, V.8, 2020, pp.23022-23040.
[37] Shan, S., Yang, P., Chen, X., et al. AdaBoost Gabor Fisher classifier for face recognition, *International Workshop on Analysis and Modeling of Faces and Gestures*, Springer, Berlin, Heidelberg, 2005, pp.279-292.
[38] Song, H.A., Lee, S.Y. Hierarchical representation using NMF, *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, 2013, pp.466-473.
[39] Sun, Y., Chen, Y., Wang, X., et al. Deep learning face representation by joint identification-verification, *Advances in Neural Information Processing Systems*, V.27, 2014, pp.1-9.
[40] Sun, Y., Wang, X., Tang, X. Hybrid deep learning for face verification, *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp.1489-1496.
[41] Tao, F., Zhang, H., Liu, A., et al. Digital twin in industry: State-of-the-art, *IEEE Transactions on Industrial Informatics*, V.15, N.4, 2018, pp.2405-2415.
[42] Tasa, U.B., Görgülü, T. Meta-art: Art of the 3-D user-created virtual worlds, *Digital Creativity*, V.21, N.2, 2010, pp.100-111.
[43] Tenenbaum, J.B, Silva, V., Langford, J.C. A global geometric framework for nonlinear dimensionality reduction, *Science*, V.290, N.5500, 2000, pp.2319-2323.
[44] Turk, M., Pentland, A., Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, V.3, N.1, 1991, pp.71-86.
[45] Wang, W., Han, X., Ding, W., et al. A Fermatean fuzzy Fine-Kinney for occupational risk evaluation using extensible MARCOS with prospect theory, *Engineering Applications of Artificial Intelligence*, V.117, 2023, pp.105518.

[46] Wright, J., Yang, A.Y., Ganesh, A., et al. Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, V.31, N.2, 2008, pp.210-227.
[47] Xu, F., Wang, H., A discriminative target equation-based face recognition method for teaching attendance, *Advances in Mathematical Physics*, V.2021, 2021, pp.1-11.
[48] Zhao, W., Chellappa, R., Phillips, P.J., et al. Face recognition: A literature survey, *ACM Computing Surveys (CSUR)*, V.35, N.4, 2003, pp.399-458.
[49] Zheng, T., Deng, W., Hu, J. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments, *arXiv preprint arXiv:1708.08197*, 2017.
[50] Zhong, J., Zheng, Y. Empowering future education: Learning in the Edu-Metaverse, *International Symposium on Educational Technology (ISET)*, IEEE, 2022, pp.292-295.



Hong Zhang - has worked in the School of Computer Science and Technology of Beihua University, since 2003. He received his M.Sc. degree in Computer Technology field Engineering from Changchun University of Technology, P.R. China in 2009. He received the B.Eng. degree from Shenyang Ligong University. His research interests include big data technology, artificial intelligence, and teaching management.



Yanan Sun - received her M.Sc. degree in Computer Science and Technology from Jilin University, China in 2005. Since 2005, she has worked in the School of Computer Science and Technology of Beihua University. Her research interests include database technology, big data and artificial intelligence.



Adam Sowik (Senior Member, IEEE) - received the B.Sc. and M.Sc. degrees in Computer Engineering and Electronics, in 2001, and the Ph.D. degree in Electronics with distinction from the Department of Electronics and Computer Science, Koszalin University of Technology, Koszalin, Poland, in 2007, and the Dr. Habil. degree in Computer Science (Intelligent Systems) from the Department of Mechanical Engineering and Computer Science, Czestochowa University of Technology, Czestochowa, Poland, in 2013. Since October 2013, he has been an Associate Professor with the Department of Electronics and Computer Science, Koszalin University of Technology.
His research interests include soft computing, computational intelligence and particularly, bio-inspired optimization algorithms and their engineering applications.



Lingling Zhang - received the B.Eng. degree in Computer Science and Technology from Beihua University, China in 2003. Since 2003, she has worked in the School of Computer Science and Technology of Beihua University. She received her M.Sc. degree in Computer Technology field Engineering from Changchun University of Technology, China in 2009. In 2015, she was rated as an associate professor. Her research interests include software engineering, big data technology and artificial intelligence.



Jianhui Lv - received B.Sc. degree in Mathematics and Applied Mathematics from the Jilin Institute of Chemical Technology, Jilin, China in 2012, M.Sc. and Ph.D degrees in Computer Science from the Northeastern University, Shenyang, China in 2014, 2017 respectively. He is currently an Associate Professor at the Pengcheng Lab., China. His research interests include artificial intelligence, ICN, in-network caching enabled networks, IoT, bio-inspired networking, evolution computation, edge computing, smart city, etc.