Establishment and Optimization of Video Analysis System in Metaverse Environment

Dandan WANG, Tianci Zhang
School of Digital Media and Art Design
Nanyang Institute of Technology
Nanyang 473000, China

***Abstract*—The current source space communication architecture has not changed. At present, the key technology of so-called metaverse media only applies its elements to existing communication architectures, and more importantly, this type of integration involves individual examples and promotional marketing methods. How to become a new growth point for the deep collaborative development of metaverse media requires strengthened research and exploration. Although AI analysis technology is powerful, its sensitivity, accuracy and adaptability could be more satisfactory due to the complexity of real-world scenarios. Given the shortcomings of existing research, we designed a video analysis system in the metaverse environment, combining virtual reality and artificial intelligence, with video perception, network, and information technology as the medium and big data as the technical support to build a full intelligence Video analysis system. The system is based on the YOLOv3 model, combined with the actual video scene, and analyzes according to the video's human behavior and environmental changes. Experiments show that the system has obvious advantages in the accuracy and recall rate of video analysis and detection, the system detection performance is significantly improved, and the video target analysis and detection of complex scenes are realized.**

***Keywords*—Artificial intelligence; metaverse; video perception; big data**

I. INTRODUCTION

In the 1960s, Ivan Sutherland et al. proposed a method to interact in the virtual world and gain experience, which can have a more real human-computer interaction experience [1], [2]. Subsequently, the first virtual reality system was developed. In the 1980s, James David Foley pointed out three elements of virtual reality: interaction, behavior and imaging. He proposed that the next generation of computers will have the complex computing power to realize the communication between users and computers and put forward the concepts of virtual reality, human-computer interaction and so on [3]. Virtual reality technology simulates the human perception function in the simulated computer environment, provides real interactive feedback, and allows people to have an immersive experience. Virtual reality affects the object form in the physical world by its real interactive feeling and good user experience so that users can gain experience in the virtual environment. Virtual reality technology can simulate the real video scene and detect the target of the real-time monitoring video, which can be adapted to the target detection task under various bad weather and low light conditions. The real-time video data can be transmitted to the virtual platform to view real-time monitoring in the virtual scene. In 2016, the virtual reality technology innovation and industrial development conference was held so that virtual reality technology has a wider range of applications in 3D printing [4], remote sensing mapping, smart cities, digital museums, teaching and training, game animation, aerospace and other fields.

In 2021, based on virtual reality technology, the concept of "metaverse" was mentioned again. Metaverse integrates virtual reality technology, creates a virtual digital world parallel to the real world through exclusive equipment, and creates a social platform with a strong sense of immersion. Users can produce and live in a virtual environment different from the real world and interact in an immersive metaverse environment. Metaverse includes the following features: user participation in creation, providing exploration space, a fusion of the virtual world and the real world, a fusion of time and space, a fusion of three-dimensional technology and intelligent technology, etc. The primary implementation method of metauniverse is virtual reality technology. VR has the advantages of creating realistic scenes, visualizing abstract content, and creating immersive experiences. Artificial intelligence (AI) technology includes deep learning, natural language processing, etc. [5]. Through the use of AI technology, we can increase the depth of research and improve the accuracy and accuracy of video analysis.

The current application scenarios that can only be monitored are becoming increasingly complex, and its classic abnormal behavior analysis rules can only be applied to some industries in China. The current algorithm rules are difficult to meet in all behavior analysis scenarios. Therefore, many applications are in their early stages and need improvement. The main factors hindering the development of intelligent video surveillance and detection systems are:

1) Environmental factors have a significant impact on intelligent video surveillance and detection systems. At present, most video analysis systems are developed based on traditional computer vision technology and do not provide image enhancement functions such as improving video image quality and clarity.

2) Many video object analysis lacks objective experimental support. Excessive reliance on reasoning leads to subjective assumptions in the interpretation of unconscious psychological phenomena.

3) There is a problem of insufficient monitoring accuracy in the experiment. The accuracy of integrated virtual reality technology in video surveillance is not satisfactory.

In the current metaverse environment, the deployment of video monitoring systems plays an important role in preventing and stopping crime, maintaining social and economic stability, and protecting the safety of the life and property of the country and people. For example, video analysis is used in 1300 channels of the Qinghai Tibet railway to comprehensively and effectively protect the safety of the whole railway. The main trend of future development of video analysis is to use deep learning technology to improve detection efficiency and realize real-time and pre-intelligent analysis. Therefore, intelligent video surveillance has become one of the most cutting-edge development fields. High definition, networking and intellectualization are the main development trends of video monitoring technology. Intellectualization refers to using intelligent analysis algorithms to process video data, extract valuable information, and prompt users with key information in the form of alarm, to improve the intellectualization and actual use value of video monitoring systems [6]. Intelligent video analysis technology can transform video monitoring from traditional "passive monitoring" to intelligent "active monitoring" and can free users from monotonous and boring monitoring work, avoid the problem of attention loss caused by long monitoring time, and realize 24-hour monitoring, so it has attracted more and more users' attention.

The research and innovation contributions are as follows:

1) The metaverse can integrate virtual reality technology and coexist. This means that the research and development of metaverse media is not only reflected in game types, but mainly in social media full scene CNC.

2) This article constructs a fully intelligent video analysis system. The system is based on the YOLOv3 model, combined with actual video scenes, and analyzed based on human behavior and environmental changes in the video.

3) YOLO3 draws on the residual network structure to form deeper network layers and multi-scale detection, improving the performance of mAP and small object detection.

4) The system has improved prediction accuracy while maintaining its speed advantage, especially enhancing its ability to recognize small objects.

This article analyzes video analysis and monitoring issues in the metaverse environment. The video analysis system in the metaverse environment has the advantages of high real-time accuracy and strong robustness, fully meeting production needs. The first part introduces the current status of the metaverse environment and video analysis technology. Section II analyzes the problems and development prospects of existing video analysis and detection methods. Section III mainly introduces the requirement analysis, design, and implementation of an intelligent video analysis system based on the metaverse environment. Section IV provides a detailed introduction to the development and testing of the system. Section V provides a summary of the entire text. The experiment shows that the system has significant advantages in terms of accuracy and recall in video analysis and detection, and the detection performance of the system is significantly improved, achieving video object analysis and detection in complex scenes.

II. Related Work

Intelligent video analysis technology is gradually applied in digital video surveillance and has been developed for over 10 years. However, regarding the current development level and actual use of intelligent video analysis technology, smart video analysis technology has only made small-scale applications in some industries, such as banking, transportation, and justice, and more often, it is only used as a highlight demonstration of the project. The market for intelligent video analysis technology has remained the same. The main reason is a large deviation between the user's cognition and demand for smart products and the degree of intelligence that intelligent products can achieve. This problem is more prominent because some manufacturers adopt the propaganda method of exaggerating performance to occupy the market. From the perspective of the long-term development of intelligent video analysis technology, it should be based on the current technical level, improve adaptability and robustness to different scenarios, simplify product settings, improve ease of use, bring real value to customers, and truly make intelligent video Analytical techniques are widely used. For example, the original monitoring images collected at the front end are structurally analyzed through intelligent video analysis technology. The original video image data are automatically transformed into quasi-structured and structured data to form the corresponding subject database. The data are submitted to the big data platform for relevant data models, technical methods and other uses, developing rich practical applications, such as human and vehicle trajectory characterization, foothold analysis, prediction and early warning and other services, giving full play to the reasonable value of monitoring images.

The rise of artificial intelligence technology and the development of network technology have pointed out new research directions for video analysis technology. Connect the hardware device to an IP network, and the webcam can be read by any device on the network via IP access. At the same time, the connection between various devices in the network relies on the standardized TCP/IP protocol. As long as the machines that meet the requirements of the protocol can access the network, the system's scalability has been greatly enhanced. In terms of algorithm, combined with the current deep learning algorithm, a role in a neural network is constructed to analyze and recognize video targets, which greatly improves the accuracy of analysis and detection. Combined with various embedded devices, controllable management is realized through the Internet/WiFi/LAN. It is highly integrated with access control, voice system, alarm devices, etc., combined with artificial intelligence algorithms to realize automatic detection, personnel identification, identity control and other functions between various machines [7].

Abnormal behavior detection of moving targets is one of the main functions of an intelligent video analysis system. Its primary task is to determine the moving targets then realize the tracking of moving targets in video and abnormal behavior analysis [8]. The traditional method to determine the moving

target is mainly realized by motion detection methods, including (1) the inter-frame difference method, which expands each frame taken by the camera according to the time sequence. Then it makes a difference between the images of the previous and subsequent frames and then regards different places in the video frame as moving targets. The biggest advantage of the inter-frame difference method is that it is simple and fast, while the biggest disadvantage is that it is difficult to separate the complete moving target. (2) Background difference method, first establish the background model of the video, then make a difference between the video frame of the monitoring system and the background model, and then locate the moving target in the video [9]. The biggest advantage of the background difference method is that it can adapt to the complex and changeable surrounding environment, and the biggest disadvantage is that it is easily affected by light. In order to make the background difference method have a good effect, it is necessary to update the background model in time. (3) The optical flow method infers the moving direction and speed of the object according to the change law of the recognized image pixel value with time. Its disadvantage is that it requires a large amount of calculation and a high computer processing capacity.

With the in-depth study of deep learning, target detection based on neural networks has higher accuracy and better stability than traditional target detection methods. It can meet the real-time requirements of video analysis. At present, deep learning algorithms for target detection are mainly divided into two categories: the two-stage method, whose principle is to divide the whole detection process into two parts to generate candidate frames and to identify objects in structures. It mainly includes: (1) region-based convolutional neural networks (R-CNN) target detection method, which greatly improves the accuracy of target detection [10]. The principle is to obtain candidate regions through selective search, then extract the features of each candidate region using a deep convolution network and carry out support vector machines (SVM) classification, and then obtain preliminary detection results. Finally, to get a more accurate boundary box, we use the deep convolution network feature combined with the SVM regression model again. (2) Fast R-CNN, this method no longer convolutes the candidate regions but convolutes the whole image and completes the fitting of the bounding box and the classification of candidate regions through the structure of the dual task network. Compared with R-CNN, Fast R-CNN shortens the training and testing time by nearly nine times [11]. (3) Faster R-CNN, this method directly generates candidate regions through the convolutional neural network, composed of regional candidate generation network and Faster R-CNN, which realizes candidate region classification and border regression. This method's training and testing time is 1/10 of that of Faster R-CNN [12]. However, the target detection method based on a convolutional neural network needs to input a fixed-size image, and the normalized production organism is truncated or stretched, which will lead to the loss of some information in the input image. The other is one stage method, whose principle is to unify the whole process and get the detection results directly, which makes the target detection process simple and efficient. It mainly includes (1) YOLO (you only look once), which directly obtains the category and location information of the target through a neural network convolution operation and realizes the real recognition and judgment in one step [13]. YOLO target detection process is very simple, and compared with the above detection algorithm, its efficiency is very high, and the processed image can reach 45 frames per second. (2) Single shot multi-box detector (SSD) [14], this method carries out detection and classification together and filters out some detected results with unclear classification and low scores. It is a detector based on a full convolution network, which can detect objects of different sizes with different layers. However, the target detection model trained by this algorithm will have the problem of positioning errors when identifying targets.

In summary, the nonlinear transformation ability of neural networks is also stronger. In target detection tasks, we need to map the input low dimensional features to a high-dimensional space in order to better represent and distinguish different targets. Traditional methods often require the design of complex mapping functions, and neural networks can learn this mapping relationship from data through automatic learning, avoiding complex design processes. Finally, the training process of neural networks can be automated and parallelized, which greatly shortens training time and improves efficiency. Of course, there are also some challenges and limitations in the target detection methods of neural networks. For example, they require a large amount of data for training and require high-performance hardware and software to run. In addition, the design and training of neural networks also require certain experience and skills, such as selecting appropriate network structures, optimizing algorithms, loss functions, etc.

III. Establishment and Optimization of Video Analysis System in Metaverse Environment

A. Traditional Video Detection and Analysis Theory

The traditional target detection system generally comprises three modules, including a region selection module, feature extraction module and classifier classification module, as shown in Fig. 1.

Fig. 1. Traditional food detection and analysis steps.

1) Region selection: The main function of the area selection module is to locate the target's location. Because the mark may appear anywhere in the image, and the target's size and length-width ratio is uncertain, the sliding window

strategy is initially used to traverse the whole picture, and different scales and length-width ratios need to be set. Because the appearance of the target is random, in the selection of the area where the target is located, the target can only be found by traversing and selecting by sliding on the original map, which leads to more redundant target boxes, which not only affects the speed, but also consumes many computing resources, and the detection results are also unsatisfactory. The higher the structure levels, the worse the system's accuracy.

2) Feature extraction: The main function of the feature extraction module is to calculate and count the image's local or global feature descriptors to get a feature map with richer semantic meaning. Affected by complex environments and scenes, in the face of diverse scene changes, traditional feature extraction methods are difficult to meet high accuracy requirements, so they can not be adapted to application scenarios. The previous feature extraction is all manually designed features, such as Haar features, hog features, etc. Still, due to the ever-changing world and limited cognitive ability, the artificially created features have various problems, such as poor robustness and unreliability, which makes it difficult to improve the efficiency of traditional target detection.

3) Classifier classification: The function of the classifier is to judge the category of a new observation sample based on the training data of the marked category. Traditional classification is mainly realized by AdaBoost and SVM [15]. Two main reasons affect classification: the structure of the classifier itself and the feature parameters extracted through the previous layer.

B. Overall Function Design of Video Analysis System in Metaverse Environment

According to the requirements of the system, five functions are realized, including the function of visual interface, the process of controlling the ball machine to track the moving target in real-time, the function of detecting the abnormal behavior of the target, the function of image enhancement and the process of panoramic stitching. The functional structure of the system is shown in Fig. 2.

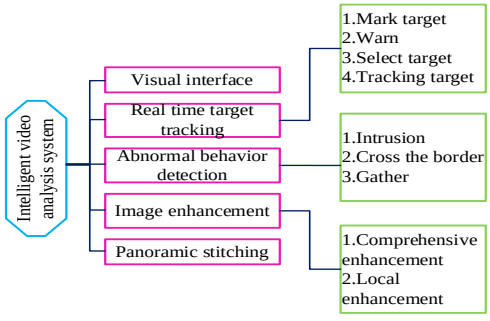

Fig. 2. System function structure diagram.

C. System Hardware Design

The main functions of the hardware system in the overall system architecture include data acquisition, video recording, result display and other parts. The data collected by different hardware is transmitted and summarized into the system software through network communication, and then the collected information is analyzed and processed through the system software. Finally, the results are stored and displayed. For the whole system, the hardware system is the premise to ensure the stability of information collection and the comprehensiveness of information collection. A reasonable hardware system architecture can ensure the stability of information collection. The correct hardware installation method and the appropriate hardware layout position are the prerequisites for comprehensive information collection. The hardware system structure is shown in Fig. 3. In this study, the hardware architecture is constructed using a network camera, network HD hard disk video recorder, monitoring hard disk, display, computer, switch and other hardware.

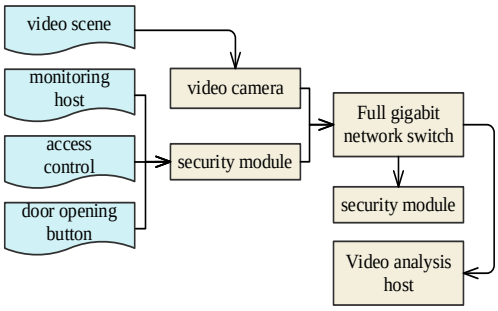

Fig. 3. Hardware system structure diagram.

D. System Software Design

The software of the intelligent video analysis system for regional control is mainly composed of four parts: data acquisition module, image processing module, database storage and processing module and result display module. First, the system obtains real-time video information and access control information through the data acquisition module, and then the image processing module performs real-time detection and processing for the monitoring area. The database storage and processing module stores and processes the results of the image processing module and the entry and exit records of the access control all-in-one machine. The result display module displays the real-time processing results of the image processing module, and the database on the interface queries the results or exports them as CSV files. The software flow chart of the intelligent video analysis system is shown in Fig. 4.

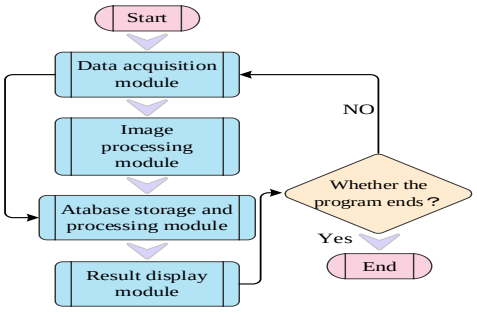

Fig. 4. Flow chart of the video analysis system.

1) Data acquisition module: There are two kinds of hardware devices used in the data acquisition module, one is

the surveillance camera, and the other is the integrated access control machine. Among them, the data acquisition of the access control all-in-one machine uses the alarm callback function in its SDK. When someone opens the door through the access control, the alarm callback function is triggered to obtain the personnel job number, access control machine IP, travel time and other related information. The information acquisition of the surveillance camera needs to cooperate with the image processing module. The data acquisition completes the equipment login, decoding, format conversion and other operations in the main thread, and the image processing module processes the converted image in the sub-thread. The multithreading processing method used in this system can not only strengthen the robustness of the system operation but also speed up the system processing speed.

First, the system needs to initialize the device SDK to prepare for the subsequent program operation. Then, the system reads the login device information. Due to different devices, the login and following processing processes are separate. The access control all-in-one machine does not need to maintain a link with the system for data transmission at all times after successful login. Only when someone triggers the alarm callback function when passing through the access control will it return a message containing the device IP to the system? It occupies relatively few computing and memory resources through the personnel job number and passing time, so its data acquisition process is in the same thread. However, the surveillance camera's image information needs to carry out real-time data transmission with the system, so it is necessary to set the disconnection reconnection time and an abnormal callback function to ensure the system's stable operation. After the surveillance camera logs in, each camera will return a device ID number, which will be used to identify subsequent code stream data differentiation of different cameras. After the camera logs in, you need to open the playback library and call the real-time stream callback to set the parameters of the video stream. Then, start the decoding callback to decode the video stream into UV12 format and convert UV12 format into RGB format for subsequent image processing.

Opening up threads and caches for each camera can first make the system process the images of each camera without interference, improve the system's stability, improve the system's computing efficiency, and enhance the real-time performance of the system. Collect image information for each camera's image cache in the system's main processing thread, process the image data in the corresponding sub-processing thread, and empty the processed image cache to keep the video frame cache updated in real-time.

2) Image processing module: The image processing module of this system is mainly used in the sub-thread of each camera of the system. The image processing module is the core of this system, which needs to combine the detection algorithm based on a convolutional neural network with the database to complete the work. Among them, the detection algorithm based on a convolutional neural network is used to detect intruders in the set deployment scenario. The database part is responsible for analyzing the time when the intruders are detected and recording the start time and end time of each intrusion. After obtaining the current frame from the video frame buffer, the image processing module uses the YOLOv3 detection algorithm after fine-tuning the scene to detect the staff. When personnel are detected, the current detection time is determined. If the current detection time is within the preset deployment time, the current detection time is compared with the intrusion time in the database. The design idea of this module is that if there is a repeated intrusion within three seconds, it will be regarded as the same intrusion event. If the interval between the latest and last detected intrusion is greater than three seconds, it will be considered a new intrusion. Therefore, the current detection time is first compared with the latest intrusion end time. If the time difference is less than three seconds, the newest intrusion end time is updated to the current detection time. If the time difference exceeds three seconds, it must be compared with the latest intrusion start time. If the time difference exceeds three seconds, the current time is recorded as the new intrusion start time. Otherwise, it is recorded as the new intrusion end time. After completing the above operations, clear the current frame data, obtain and process the next frame.

3) Database storage and processing module: This system's database storage and processing module is mainly used to process the access information returned by the access control machine. Through access control information and face comparison based on video frame images, the system can realize several functions, such as attendance management, overtime detection of non-secret related personnel in hidden related areas, detection of non-secret related personnel entering illegally, and reminder of VIP visits. The four functions of this module are realized by hardware information triggering the query and processing of the database.

a) Attendance management: First, the real-time access information returned by the access control is written into the database, and the database information of the previous day is processed at 0:00 every day. Each person's first access is recorded as the time of work, and the time of the last entry is recorded as the time of departure. Then, they are compared with the preset on-off time, the day of the week and other information to obtain whether there are late, early leave, absenteeism or overtime on that day, and store the analysis results.

b) Overtime detection of non-secret related personnel in hidden related areas. The deployment site of this system requires that when non-secret-related personnel enter the secret-related room, whether the door is locked or not, they must pass the access control and stays for no longer than the specified time. Therefore, when the system detects that non-classified personnel enter the classified area after a specified time of waiting, it will query the access information of the classified site during the waiting period. If the non-classified person has no record of entering and leaving the classified site within this period, the non-classified person will be recorded as an overtime stay.

c) Detection of illegal entry of non-secret related personnel: When the camera at the entrance of the classified area detects that a non-classified person has entered the classified area through face comparison, it determines whether the non-classified person has passed the access control when entering the classified area by querying and analyzing the database. If the result is that he has yet to enter the classified area through access control, record the illegal entry behavior.

d) Important guest visit reminder: When the camera outside the main door detects the person through face comparison, it determines whether the visitor is a VIP by querying the VIP information in the database. If it is a VIP, it will send a reminder message.

e) Result display module: The results show that the module's structure is relatively simple and consists of two parts. One part displays the real-time detection effect of the image processing module on the software interface, and the other part displays the data queried from the database on the interface in the form of a list and exports it to a CSV format file. This module only displays image data and database data without processing.

E. Key Technologies of the Video Analysis System

1) Back propagation algorithm: The backpropagation algorithm is important for convolutional neural networks to calculate gradients. Dr Paul proposed the backpropagation algorithm werbos in his doctoral thesis in 1974, but it was only widely recognized once David proposed it again in 1986. At present, this is one of the most commonly used algorithms for training neural networks. The input of the neural network structure is a two-dimensional vector (x_1, x_2), and the corresponding parameters to be optimized are w_1, w_2 and b. The last step of the nonlinear transformation operation is completed by a ReLu [16]. The calculation is divided into the most basic multiplication and addition operations. When backpropagation is required, multiply all the gradients on the dotted line from y to the leaf node.

At present, the most widely used function is the ReLu function. Its linear and unsaturated characteristics give it strong convergence ability in model training.

$$y = \begin{cases} 0 & (x \le 0) \\ x & (x > 0) \end{cases} \quad (1)$$

In the ordinary ReLu function, for the input less than 0, the derivative is constant to 0, which may cause too many silent neurons. Leaky ReLu [17], as an improved version of the ordinary ReLu, avoids this problem by introducing a normal numberλpredefined to be close to 0.

$$y = \begin{cases} x & (x > 0) \\ \lambda x & (x \le 0) \end{cases} \quad (2)$$

Compared with the ReLu function, the PReLu function introduces value α, which is obtained by adaptive learning based on data. The introduction of a can prevent the gradient from disappearing, accelerate the convergence speed and reduce the error rate.

$$y = \begin{cases} x & (x > 0) \\ \alpha x & (x \le 0) \end{cases} \quad (3)$$

Leaky ReLu and PReLu have the best relative performance based on the above common activation functions. Their linear and unsaturated characteristics can help the model converge quickly. At the same time, they both solve the problem of training termination caused by silent neurons caused by gradient disappearance. However, parameter a in PReLu needs adaptive learning, and the YOLOv3 algorithm is relatively sensitive to speed, so it is necessary to reduce the number of learning parameters so the activation function selected in YOLOv3 is Leaky ReLu.

2) Gradient descent algorithm: When YOLOv3 uses the back-propagation algorithm to calculate the gradient, it also uses the gradient descent algorithm [18]. Generally, there are three kinds of gradient descent algorithms: full batch gradient descent, random gradient descent (SGD) and small batch data gradient descent. The full data gradient descent algorithm is to calculate all the training sample data every time the gradient is calculated. If the total number of samples in the training set is N, calculate the entire loss function by calculating the loss function for all N sample data and then calculate the mathematical expectation. The calculation formula is shown in Eq. (4).

$$l(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(\theta; xi, yi) \quad (4)$$

In Eq. (4), N is the total amount of training sample set, (x_i, y_i) represents a set of random components, and θ is the parameter to adjust the component weight.

The distance measurement formula used in K-means clustering is shown in Eq. (5).

$$d(box, centroid) = -IOU(box, centroid) \quad (5)$$

In Eq. (5), d represents the distance from the centroid to the bounding box, (box, $centroid$) is the geodesic distance from the centroid to the bounding box. IOU (Intersection over Union) is a measure of the accuracy of detecting corresponding objects in a specific dataset.

The second is to improve the problem of unstable model fitting in the traditional anchor algorithm. The instability in the conventional anchor algorithm is mainly caused by predicting the coordinates (x, y) of the bounding boxes. The calculation formula is shown in Eq. (6).

$$x = (tx * wa) - xa$$
$$y = (ty * ha) - ya \quad (6)$$

In Eq. (6), (x_a, y_a) represent the coordinates of the center of the bounding box, t_x is the x coordinate movement step size, t_y is the y coordinate movement step size, w_a is the direction parameter of the x coordinate, h_a is the direction parameter of the y coordinate. In Eq. (6), no constraints are imposed on the bounding boxes, and any bounding box can fall on any area of the image, which is relatively difficult to stabilize. In YOLOv3, the improved anchor algorithm directly predicts the relative

position of the bounding box relative to the grid unit, and the calculation formula is shown in Eq. (7).

$$\begin{aligned} bx &= \sigma(tx) + cx \\ by &= \sigma(ty) + cy \\ bw &= pwe^{tw} \\ bh &= phe^{th} \end{aligned} \tag{7}$$

In Eq. (7), (c_x, c_y) is the coordinate of the upper left corner of the grid, p_w and p_h are the width and height of the anchor box, respectively, and (b_x,b) is the central coordinate of the bounding box, and b_w and b_h are the width and height of the bounding box. After the above constraints, the network becomes more stable.

Compared with the YOLO algorithm that does not use the anchor idea, the map has a slight decrease, but the recall rate has a significant increase, as shown in Table I. Where AP represents the average accuracy of target detection, the map represents the average AP value, and recall represents the ratio of correctly identified targets to the number of all flying targets in the test set. This is because each graph can only give dozens of prediction boxes before using the anchor idea. After using the anchor idea, the model can have more than 1000 prediction boxes, giving the algorithm more room for improvement.

TABLE I. COMPARISON OF RESULTS WITH AND WITHOUT ANCHOR

	Map	recall
Without anchor	69.5	81%
With anchor	69.2	88%

3) Loss function: In YOLOv3, the key prediction information in the final result mainly includes four categories: prediction frame length and width (w, h), coordinates (x, y), category class and confidence. According to these four information characteristics, the appropriate loss function is selected, and then the weighted average is obtained to obtain the final loss function. Among them, the loss function used in the prediction frame length and width (w, h) is the sum of squares of the errors of the corresponding points of the original data and the fitting data, and its calculation formula is shown in Eq. (8). The closer the SSE result is to 0, that is, $y_i = \hat{y}_i$, it means that the predicted parameters are equal to the real value, the network reaches the convergence state, and the higher the target detection rate, the stronger the fitting ability of the final model.

$$SSE = \sum_{i=1}^{n} wi(yi - \hat{y}i)^2 \tag{8}$$

The loss function used for the remaining three types of information is the cross entropy loss function binary_crossentropy; the calculation Formula is shown in Eq. (9) and Eq. (10).

$$loss = \sum_{i=1}^{n} \hat{y}i \; log \; y \, i + (1 - \hat{y}i) \; log(-\hat{y}i) \tag{9}$$

$$\frac{\partial loss}{\partial y} = -\sum_{i=1}^{n} \frac{\hat{y}i}{yi} - \frac{\hat{y}i}{yi} \tag{10}$$

Eq. (10) shows that only when $y_i = \hat{y}_i$, the loss value is 0, and in other cases, the loss is constant as a positive number and the greater the difference between *you* and the greater the loss value.

IV. RESULT AND ANALYSIS

According to the video scene's personnel management and control requirements, a detection and analysis scheme suitable for this scene is designed. YOLOv3 is fine-tuned on the video scene data set, and the accuracy of the fine-tuned algorithm is verified through experiments. Finally, the stability of the system is verified.

The CPU of the hardware platform used in this study is i5 4590, and the main frequency is 3.3GH; The GPU is NVIDIA's Teska K40, and the video memory is 12GB; Memory is 16GB; The system is a 64-bit ubuntu16.04 high-performance host. It needs to be done on Darknet. Before installing Darknet, you must first configure CUDA and OpenCV, recompile Darknet on the configured host, and finally get the Darknet framework loaded with CUDA and OpenCV.

After completing the above preparations, set the response path in the program and start the training. After training, a fine-tuned YOLOv3 model with a size of 246.3mb is obtained. The loss image obtained from training is shown in Fig. 5. The loss function is a function to measure the detection performance of the algorithm. In Fig. 5, at the beginning of training, the loss value decreased rapidly, indicating that the fitting speed of the model quickly increased. With the progress of the training process, the decline speed of loss value gradually slows down, indicating that the model fitting is slowly completed. When the training times reach 40000, the loss value is the minimum. In the 40000 and 45000 training, the learning rate is adjusted to the current 1/10, and the loss value tends to be stable and minimum. At this time, the training is completed, and the model reaches the optimal state.

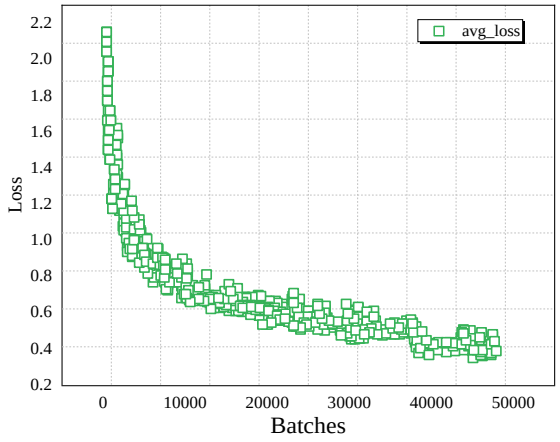

Fig. 5. Loss function training curve.

Five hundred pictures are cut from the live video images of the sample set as the test sample set. The selected pictures cover as many scenes as possible, and the proportion conforms to the actual application scene. The final test sample focuses on 300 pictures of outdoor areas under natural light during the day, with a total of 1924 personnel targets; 100 pictures of

indoor areas under natural light during the day, with a total of 563 personnel targets. There are 100 pictures in the street backlight scene, with a total of 125 personnel targets. Fig. 6 shows a group of image comparisons under the same angle, natural light in the day and light at night. The performance of system target detection under different light can be studied through the comparison test.

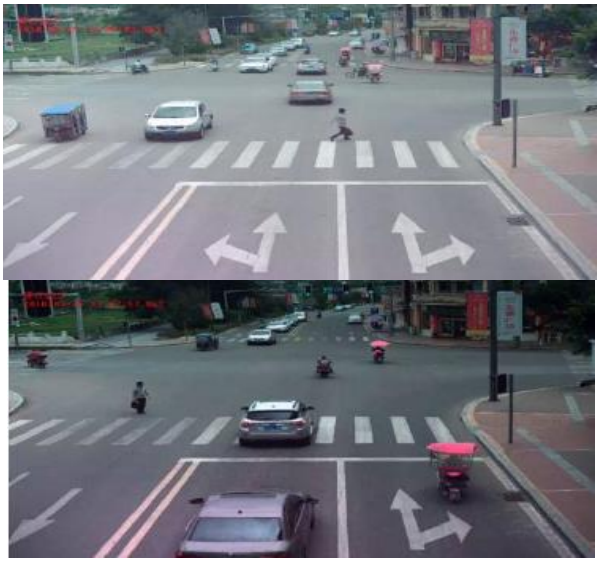

Fig. 6. Training set samples.

There are 2612 personnel targets in the test data, of which 2518 targets were correctly detected, 94 targets were missed, and the background was detected as personnel targets eight times, with accuracy P=0.997, recall rate R=0.964, false alarm rate FA=0.003, and missing alarm rate FNR=0.036. Under the condition of natural light in the daytime, the accuracy P=0.997, the recall rate R=0.960, the false alarm rate FA=0.003, and the missed alarm rate FNR=0.040; Under the condition of night light, the accuracy P=0.996, recall rate R=0.973, false alarm rate FA=0.004, missing alarm rate FNR=0.027; Under the state of the strong backlight at the sightseeing elevator entrance, the accuracy P=1, the recall rate R=0.984, the false alarm rate FA=0, and the missing alarm rate FNR=0.016. The specific test results are shown in Table II.

TABLE II. STATISTICAL TABLE OF OBJECT DETECTION RESULTS OF YOLOV3 MODEL PERSONNEL

	TP	FN	FP	R	P	FNR	FA
Daytime indoor conditions	1847	77	6	0.960	0.997	0.040	0.003
Daytime outdoor conditions	548	15	2	0.973	0.996	0.027	0.004
Backlight condition	123	2	0	0.984	1	0.016	0
total	2518	94	8	0.964	0.997	0.036	0.003

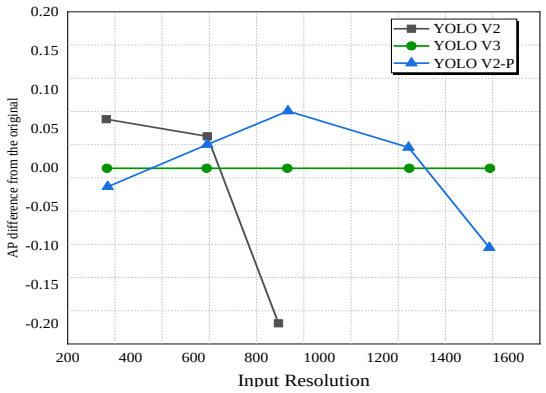

Fig. 7. AP comparison of networks under different resolutions.

As shown in Fig. 7, in the experiment, images of different resolutions are selected to input the network, and the AP of the network is tested. We can find that YOLOv3 has the best AP at high resolution, while YOLOv2 and YOLOv2-p have the best AP at middle and low key, respectively. This means that we can use sub-nets of YOLOv3 models to execute the object detection task well. Moreover, we can perform a compound scale-down of the model architectures and input size of an object detector to get the best performance.

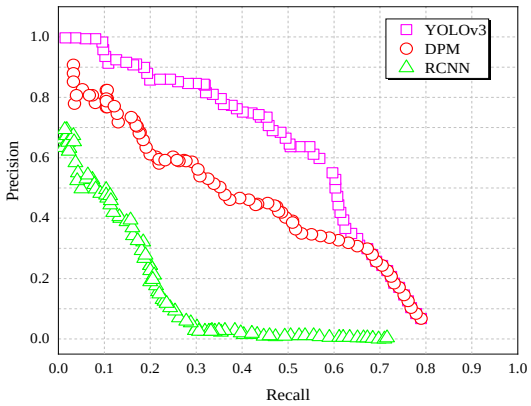

Fig. 8. Dataset precision-recall curves.

The experiment compares the traditional deformable parts model (DPM) target detection algorithm and the target detection algorithm based on RCNN with this method for human detection in the data set. As shown in Fig. 8, it can be seen that the YOLOv3 model still has a high correct detection rate when the recall rate reaches 60%, but the detection rate of the DPM algorithm and the RCNN method decreases significantly. Because the DPM target detection method is characterized by manual marking, it has certain limitations, which leads to performance degradation. After transforming the detection problem into the classification problem of the local area of the picture, the RCNN method cannot fully use the context information of the local features of the image in the whole picture and loses the global attributes, resulting in performance degradation.

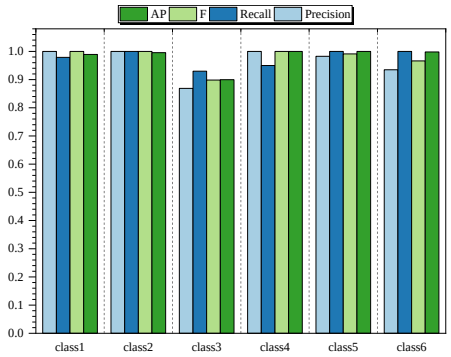

Fig. 9. Object classification detection accuracy comparison.

As shown in Fig. 9, the experiment selects six types of detection targets in the data set for classification and detection and compares the detection accuracy. It can be seen that the YOLOv3 network model maintains high detection accuracy for six categories of targets. Category three is a selected low-resolution group, contrasting categories with light interference, and the detection accuracy is slightly reduced. Fig. 10 is a simulation of the classification and prediction results of the detection targets by the network; the target classification method is based on the target detection technology to locate the targets in the image, analyze the characteristics, and cluster the targets with the same features to form a certain category. It can be seen from the figure that the network has more accurately classified and identified the detection targets of the two categories.

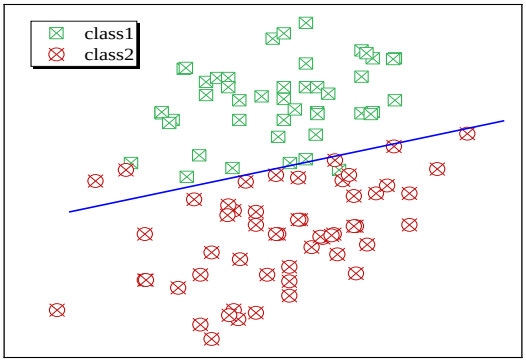

Fig. 10. Object classification prediction simulation.

V. Discussion

Obviously, the root filter model describes the overall characteristics of a person, and if only one model is used to detect objects, it is definitely not as effective as detecting multiple models. So DPM also has a component filter model, and the number and parts of the component filters can be designed by oneself. The component filters describe the local features of a person's head, hands, and feet.

The local feature position detected by the component filter must not be too far from the position of this local feature in the root filter. Imagine if the distance from the hand to the body is twice the height, then is this still a person. So DMP added the position offset between the component filter model and the root filter model as the offset coefficient. As shown in the spatial model in the figure, the center of each box naturally represents the rational location of the component model. If the position of the detected component model happens to be here, the offset coefficient will be 0, and a certain value will need to be subtracted from the surrounding area. The further the deviation, the greater the value will be subtracted. That is to say, subtracting the offset coefficient from the comprehensive score essentially involves using the spatial prior knowledge of the root model and component model.

Compared with the traditional anchor idea, YOLOv3 mainly makes two improvements. First, the method of manually selecting bounding boxes in the conventional anchor idea is changed, and a more reasonable K-means clustering method is selected. Compared with the manual selection of bounding boxes, some tall, thin and short boxes appear in the boxes obtained after K-means clustering [19]. Although unsupervised feature selection can remove some irrelevant and redundant features in some cases, it is usually more difficult than supervised feature selection because it does not have class information to help determine which features are relevant [20]. During detection, the inner product of the DMP feature vector of the input image and the filter operator is calculated to obtain the response value of this filter operator. After comprehensive different filtration the response value of the wave operator can be calculated as a comprehensive score, and then trained as a score threshold to detect humans.

VI. Conclusions

With the change and development of society, information technology has had a wide and profound impact on today's society. Making full use of modern information technology has become an objective trend of social development; as one of the overall goals of information construction, current personnel area monitoring and management system needs to combine advanced digital image processing technology, advanced computer information technology and exchange network technology to meet the needs of high reliability, stability, security and applicability. This study constructs an intelligent video analysis system based on the metaverse environment. The procedure takes the YOLOv3 model as the core and fine-tunes it to enhance its detection ability. Experiments show that the fine-tuned YOLOv3 model has been significantly improved.

From the perspective of model structure, the YOLOv3 model does not use pooling layers and fully connected layers, but instead sets the convolutional stripe to two to achieve downsampling. Every time this convolutional layer is passed, the size of the image will be reduced to half. However, mainstream systems such as Windows, MacOS, Linux, Android, iOS, and Chrome OS do not have such model structure characteristics. Secondly, in terms of setting prior boxes, YOLOv3's method uses K-means clustering to obtain the size of prior boxes. On the COCO dataset, YOLOv3 clustered a total of 9 sizes of prior boxes. These prior boxes are set based on different scales of image features, enabling the model to better adapt to target detection at different scales. However, this method is not adopted by mainstream systems. Combined with the development of intelligent video surveillance, pedestrian detection technology and deep learning at home and abroad, this research constructs a smart personnel

area management ad control system that meets the actual needs. The construction scheme covers both hardware and software. This study selects the hardware equipment suitable for this system and designs an appropriate layout scheme according to the actual situation on site. In terms of software, this study combines a data acquisition module, image management module, database storage and processing module and result display module to complete various intelligent management requirements. Although this research realizes the simultaneous detection of multiple video streams, the detection results of each video stream need to be more connected. In the future, we can combine the detection results of each video stream by using technologies such as Reid to form a more powerful system.

Competing of interests: The authors declare no competing of interests.

Authorship Contribution Statement: Dandan Wang: Writing-Original draft preparation, Conceptualization, Supervision, and Project administration. Tianci Zhang: Methodology, Language review, and Validation.

REFERENCES

[1] G. B. Petersen, G. Petkakis, and G. Makransky, "A study of how immersion and interactivity drive VR learning," Comput Educ, vol. 179, p. 104429, 2022.

[2] H. Wu, T. Cai, D. Luo, Y. Liu, and Z. Zhang, "Immersive virtual reality news: A study of user experience and media effects," Int J Hum Comput Stud, vol. 147, p. 102576, 2021.

[3] J. Brade, M. Lorenz, M. Busch, N. Hammer, M. Tscheligi, and P. Klimant, "Being there again–Presence in real and virtual environments and its relation to usability and user experience using a mobile navigation task," Int J Hum Comput Stud, vol. 101, pp. 76–87, 2017.

[4] M. Davia-Aracil, J. J. Hinojo-Pérez, A. Jimeno-Morenilla, and H. Mora-Mora, "3D printing of functional anatomical insoles," Comput Ind, vol. 95, pp. 38–53, 2018.

[5] T. Kempitiya, S. Sierla, D. De Silva, M. Yli-Ojanperä, D. Alahakoon, and V. Vyatkin, "An Artificial Intelligence framework for bidding optimization with uncertainty in multiple frequency reserve markets," Appl Energy, vol. 280, p. 115918, 2020.

[6] K. Kanai, K. Ogawa, M. Takeuchi, J. Katto, and T. Tsuda, "Intelligent video surveillance system based on event detection and rate adaptation by using multiple sensors," IEICE Transactions on Communications, vol. 101, no. 3, pp. 688–697, 2018.

[7] C. Zhao et al., "Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images," Pattern Recognit, vol. 119, p. 108071, 2021.

[8] H.-M. Hsu, J. Cai, Y. Wang, J.-N. Hwang, and K.-J. Kim, "Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model," IEEE Transactions on Image Processing, vol. 30, pp. 5198–5210, 2021.

[9] M. Uzair, R. S. A. Brinkworth, and A. Finn, "Bio-inspired video enhancement for small moving target detection," IEEE Transactions on Image Processing, vol. 30, pp. 1232–1244, 2020.

[10] Z. Zhao, X. Li, H. Liu, and C. Xu, "Improved target detection algorithm based on libra R-CNN," IEEE Access, vol. 8, pp. 114044–114056, 2020.

[11] S.-S. Baek et al., "Identification and enumeration of cyanobacteria species using a deep neural network," Ecol Indic, vol. 115, p. 106395, 2020.

[12] L. Zeng, B. Sun, and D. Zhu, "Underwater target detection based on Faster R-CNN and adversarial occlusion network," Eng Appl Artif Intell, vol. 100, p. 104190, 2021.

[13] H. Zhou, Y. Zhao, and W. Xiang, "Method for judging parking status based on yolov2 target detection algorithm," Procedia Comput Sci, vol. 199, pp. 1355–1362, 2022.

[14] L. Cai, F. Dong, K. Chen, K. Yu, W. Qu, and J. Jiang, "An FPGA based heterogeneous accelerator for single shot MultiBox detector (SSD)," in 2020 IEEE 15th International Conference on Solid-State & Integrated Circuit Technology (ICSICT), IEEE, 2020, pp. 1–3.

[15] A. Gavriilidis, J. Velten, S. Tilgner, and A. Kummert, "Machine learning for people detection in guidance functionality of enabling health applications by means of cascaded SVM classifiers," J Franklin Inst, vol. 355, no. 4, pp. 2009–2021, 2018.

[16] C. Banerjee, T. Mukherjee, and E. Pasiliao Jr, "An empirical study on generalizations of the ReLU activation function," in Proceedings of the 2019 ACM Southeast Conference, 2019, pp. 164–167.

[17] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization," IEEE Transactions on Signal Processing, vol. 67, no. 9, pp. 2357–2370, 2019.

[18] D. Wu et al., "Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector," Biosyst Eng, vol. 189, pp. 150–163, 2020.

[19] Z. Feng, W. Niu, R. Zhang, S. Wang, and C. Cheng, "Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization," J Hydrol (Amst), vol. 576, pp. 229–238, 2019.

[20] H. Hu, R. Wang, X. Yang, and F. Nie, "Scalable and flexible unsupervised feature selection," Neural Comput, vol. 31, no. 3, pp. 517–537, 2019.