Research on Intelligent Campus Monitoring Management System Based on Deep Neural Network Algorithm

XiaLiu1, Xuelong Wang2 and ChangLin Ren3
1Department of Information Center, Xi'an Shiyou University, Xi'an, 710065, China
2Department of Information Center, Xi'an Shiyou University, Xi'an, 710065, China
3Department of Information Center, Xi'an Shiyou University, Xi'an, 710065, China
Corresponding author’s e-mail: 329507616@qq.com

Abstract. With the development of the economy and the advancement of science and technology, the informationization of education management has become an inevitable trend, and the construction of an intelligent campus cannot be delayed. In order to comply with the current social development, major universities have invested heavily in IoT equipment to improve the management level of the school. Based on the large number of video cameras installed in colleges and universities, this paper uses the monitoring image big data and the CNN algorithm to train a set of graphical features that are easy to manage, without adding additional hardware purchases. In addition, combined with the camera position and other information, an intelligent campus monitoring system is built to help the campus administrator to understand the people flow, traffic flow and physical condition in real time, which is convenient for enhancing the effective monitoring and management of campus crowds, environment and security.

1. Introduction

Nowadays, major universities have actively responded to the slogan of "Internet +" in order to implement the national "Education Informatization 2.0 Action Plan", and have devoted themselves to the wave of construction of the Internet of Things and big data. Universities continue to explore how to ensure campus security in a green and efficient manner; rationally manage campus areas; rationally evacuate and guide campus people; and construct a campus environment that is human, safe, and ecologically sustainable.

Campus management usually needs to know the location of what happens. Some people use GPS to locate people and things on campus, but due to the small campus area, there is a problem of low accuracy. Later, there were school cooperating with carriers by using SIM cards to locate people, but some people were reluctant to switch networks. In order to obtain more precise positioning and low energy consumption, RFID, iBeacon, NFC and other wireless technologies have also entered the research field. Such solutions require additional energy and resource inputs, and there are also limited information collection and user terminal tampering. Problems such as information are not very satisfactory. The camera as an image acquisition device, if deployed to a certain number and organized together, can form a sensor network that collects information in real time, and the event monitored by the camera is the exact location of the camera.

A large number of cameras have become commonplace, but a large number of surveillance screens have brought new problems. In recent years, in order to strengthen the security work of the school, the
security department has introduced a large number of video camera equipment, which are located in every corner of the school and observe every moment of the campus from various angles. However, the large amount of screens in the monitoring room are too many to be looked by the security guard. At present, the image data captured by video camera is used more as a post-mortem data. How to enable the system to automatically pick out the picture worth attention will greatly improve the monitoring and management efficiency.

Based on the maximization of the use of existing resources, this paper uses image recognition technology to mark existing video big data and use the convolutional neural network (CNN) [1] to train information such as people and vehicles. The feature model, combined with the campus location information of the monitoring equipment, generates a real-time campus population hotspot map, vehicle hotspot map, helps the staff to avoid the crowd, guides the vehicle, and can quickly help the security guard to identify group events, abnormal picture e.t.c. It will effectively improve the quality of security and management.

2. System architecture design

The system consists of the object-sense hardware layer (hereinafter referred to as the sensing layer), the intelligent identification software layer (hereinafter referred to as the identification layer) and the data layer. The system architecture shown in Figure 1.

![Figure 1. System architecture diagram](image)

Based on the campus map modeling, the sensing layer consists of a campus camera that captures the real-time remote sensing network. All cameras are numbered according to the shooting position. Camera registration parameters include unique ID, device model, campus model x-axis value, campus model y-axis value, campus model z-axis value (ground is 0).

The sensing layer periodically captures the screenshot image of the camera, recognizes the image information through the CNN algorithm, and performs intelligent prompting according to the type of the image information. At present, the picture information is divided into three categories: human hotspots, vehicle hotspots, and facility status. Corresponding to people event monitoring scenarios (such as group events, nighttime abnormal monitoring), vehicle guidance scenarios (such as heavy traffic, parking lot empty state), equipment status monitoring scenarios (such as whether the manhole cover is displaced, whether the street lights are damage).

The data layer provides data services for the above two layers, the camera database stores the camera information designed by the sensing layer, the campus map database stores the campus map information, the CNN recognition model database stores camera feature pattern data, and the image collection cache provides real-time image saving and a certain length of history image saving for backtracking.
3. Algorithm introduction

In the 1960s, when early researchers studied the local sensitive and directional selection of neurons in the cat's cerebral cortex, they found that their unique network structure can effectively reduce the complexity of the feedback neural network, and then proposed a convolutional neural network[2]-[4]. (Convolutional Neural Networks, CNN for short), a schematic diagram of the CNN algorithm is shown in Figure 2.

![Figure 2. CNN algorithm diagram](image_url)

CNN is mainly used to identify two-dimensional graphics of displacement, scaling and other forms of distortion invariance. CNN can implicitly learn from the training data. At the same time, because the same feature has the same weight of neurons and neural network parallel learning, the convolutional neural network has more accurate advantages in image recognition.

The CNN basic structure consists of three layers [5]. The first layer is the feature extraction layer. Each neuron scans the original image to extract features. All data after feature extraction is aggregated into a new feature layer. This process is convolution. The second layer is the feature mapping layer, which is produced the activation correction (usually using the ReLU algorithm), and the pooling process [6], reducing the image resolution of the feature map, obtaining stable features, thereby simplifying the number of features (pooling generally use the MAX or AVE method). The last layer is the fully connected layer, which fully connects all the generated feature mapping layers. All the layers are flattened to form the final feature map, and the obtained feature map has category distinction information [7].

3.1 Training data preparation

The basis for CNN recognition accuracy is accurate and appropriate training data. Therefore, for different camera scenes, we extract video data from different time periods and different seasons for random screen sampling. For the whole year, the weekly sample is taken 24 hours a day. In order to be compatible with the performance of new and old cameras, the resolution of the picture is 800*600 pixels. Then, according to the camera scene, the pictures are classified by manual classification. Finally, each of the groups of cameras is selected to have their own signature database training data. Through learning, CNN can more accurately identify the state of people, vehicles, street lamps, manhole covers and other physical objects.
3.2 Training recognition model

CNN Convolution. Since CNN has local field of view features, stride is 1, traversing each neuron (Cell), convolution with the input image to generate the first image to be processed (1st Pic), because we have to layer the image Connection processing, in order to reduce the over-fitting problem caused by multiple calculations of graphical linear features, we use the activation function to nonlinearly process the image to generate a second image to be processed (2nd Pic). As shown in Figure 3, the input layer information is convolved by neurons and then activated to generate a new feature layer.

There are many activation functions, including ReLU, Sigmoid, LeakyReLU, Maxout, and TanH. Among them, ReLU is the most commonly used and the fastest, as the first choice of this paper. LeakyReLUc and Maxout are used as alternative functions, and the rest of the functions are slow to study or not applicable. This model is not selected.

The convolution and activation process can be represented by equation (1), where x represents the gray value of the original layer's two-dimensional plane, k represents the neuron, and b represents the deviation of the activation function:

\[
x_{i,j}^{(l)} = f(u_{i,j}^{(l)}) = f(\sum_{i=0}^{n} \sum_{j=0}^{m} x_{i,j}^{(l)} k_{p,q}^{(l)} + b^{(l)})
\]

Pooling (Pooling). Due to the multi-neuron convolution process, the original input is multiplied by the number of neurons, and together with the next layer of the neural network (Hidden Layer), as a new original input. Repeatedly, it will bring huge computational pressure to the subsequent neural network processing. However, according to research by Hubel et al., the input features can be compressed to achieve the same recognition effect. Therefore, the algorithm pools the new feature data to significantly reduce the amount of feature data. The pooling process utilizes the principle of similarity in the adjacent regions of the image to compress the image. While retaining the characteristics of different neural cores as much as possible, the amount of data processed by the next layer of neural network is greatly reduced, and the feature map is obtained. Layer data (Feature).

Pooling is currently also popular with two strategies: Maximization (MAX) and Average (AVE). The maximization strategy means that the largest value in the pooled area represents the pooled area, and the average strategy calculates the average of the pooled area to represent this area. At present, there is no better conclusion. We use the maximization strategy for pooling.
After a single round of convolution processing and pooling processing, the specific process shown in Figure 4, we can get a round of feature data. By looping through this process, we can finally purify the data and concentrate the data features.

Full Connect. After several rounds of processing, we get a lot of feature image data. At this time, the traditional neural network processing flow is entered, and all the feature layer data are fully connected, so that the feature data having a considerable thickness is flattened to a plane. Finally, combined with Softmax we get classification model data.

4. The system workflow
The central system periodically get the image data of the camera from video disk, and the identification layer extracts the recognition model according to the ID to perform image recognition according to the ID information of the camera. Based on the recognition results, comprehensive monitoring results are given. If it is a flow of people information, according to different factors such as congestion and night security, give a hot traffic reminder or nighttime personnel abnormal activity warning, and give the security guard a reminder. If it is vehicle information, according to different scenes, push the information to campus traffic center to give advice. If it is maintenance facility information, give important equipment shift anti-theft prompts, or campus maintenance information such as missing manhole cover according to the situation.

5. 3D convolution in-depth application
The preliminary design of the monitoring system is based on a single-channel grayscale image to determine the image, and the data is judged as the moment when the image is generated. In this way, there is an accidental flow of people, and it is judged that the people are crowded. The root of this problem lies in judging the judgment based on the image information at a certain moment. Based on the research, this paper adds the depth to the width and height based on the RGB three-channel recognition technology of color pictures. The newly added channels represent the images of the first 5 minutes and the first 10 minutes of the image generation, so that we have a status information within ten minutes of the monitoring area, and it is judged from a transient state to a period of time. This kind of improvement has a good effect on the crowd gathering and the judgment of vehicle congestion, which is convenient for more accurately determining the occurrence of campus events.

6. Conclusion
This paper mainly uses CNN algorithm, combined with a large number of monitoring image information, to train a set of effective image recognition features. Through these identification features,
the monitoring device is helped to identify people and things, thereby forming some campus monitoring information, improving monitoring efficiency and other aspects of campus management. On the basis of this, combined with the principle of CNN to identify color pictures, a time segment pattern feature recognition method is proposed. This improvement can effectively help the system to distinguish between transient phenomena and persistence phenomena. In the end, the campus administrators were easily to identify abnormal crowd gathering events, important material theft incidents, school vehicle traffic conditions, and missing manhole covers. Realize campus security and smart management, and construct a campus environment with humanity, security and ecological sustainability.

References
[1] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86 (11) :2278-2324.
[2] PEI Zilong, XING Jinsheng. Improved Convolutional Neural Network Algorithm Based on Particle Swarm Optimization[J]. Journal of Shanxi Normal University (Natural Science Edition), 2017, 31(2):22-26.
[3] Dou Min. Design and Implementation of Video Semantic Analysis System Based on CNN and LSTM[D]. Nanjing University Of Posts And Telecommunications, 2018.
[4] Yao, Cong, et al. Strokelets: A learned multi-scale representation for scene text recognition[C]. Computer Vision and Pattern Recognition, 2014 IEEE Conference on. IEEE, 2014: 437-442.
[5] BO KangHu, LEE FeiFei, CHEN Qiu. Scene Recognition Algorithm Using Advanced CNN Features[J]. Computer Systems & Applications, 2018, 27(12):25–32.
[6] ZHOU FeiYan, JIN LinPeng, DONG Jun. Review of Convolutional Neural Network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.
[7] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:8614-8618.