Construction and Application of Indoor Video Surveillance System Based on Human Activity Recognition

Yuchen Wang[1], Mantao Wang[1,2,a], Zhouyu Tan[1], Jie Zhang[1,2], Zhiyong Li[1,2], Jiong Mu[1,2] ,Zhihao Zhou[1] and Lixing Luo[1]

[1]Sichuan Agricultural University, College of Information Engineering, 625000 Yaan, China
[2]The Lab of Agricultural Information Engineering, Sichuan Key Laboratory, 625000 Yaan, China

Abstract. With the growth of building monitoring network, increasing human resource and funds have been invested into building monitoring system. Computer vision technology has been widely used in image recognition recently, and this technology has also been gradually applied to action recognition. There are still many disadvantages of traditional monitoring system. In this paper, a human activity recognition system which based on the convolution neural network is proposed. Using the 3D convolution neural network and the transfer learning technology, the human activity recognition engine is constructed. The Spring MVC framework is used to build the server end, and the system page is designed in HBuilder. The system not only enhances efficiency and functionality of building monitoring system, but also improves the level of building safety.

1 Introduction

With the growing of monitoring network and the development of computer image processing technology, the intelligent monitoring system which includes the human activity recognition function is becoming mature gradually. It is can be applied in every aspect of the monitoring system. This way of monitoring aims to ensure safety in buildings. Nowadays, most of the buildings are equipped with video surveillance cameras in rooms and corridors. It is a great significance to prevent violence by making use of the monitoring system in the building. Combined with human activity recognition, the alarm will ring out when search out abnormal actions such as fighting and falling down. Due to the daily increasing monitoring network, human resources find it more difficult to do observation task. Moreover, the use of human activity recognition technology in monitoring the abnormal action will visibly improve the security level. This paper proposes a solution to the construction and application of indoor intelligent monitoring system, with the help of convolution neural network, human activity recognition, Spring MVC and some other technologies.

2 System functional requirements analysis

Through previous research, it is found that violence usually occurs in some corners in the building, such as the end of the corridor and so on. Through the comprehensive analysis of the factors of indoor violence and the existing security system in building, the functional modules of this system should include four parts: the display of monitoring information; the setting of monitoring; the options of the Human activity recognition engine; the system setting. The system function module diagram is shown in Figure 1.

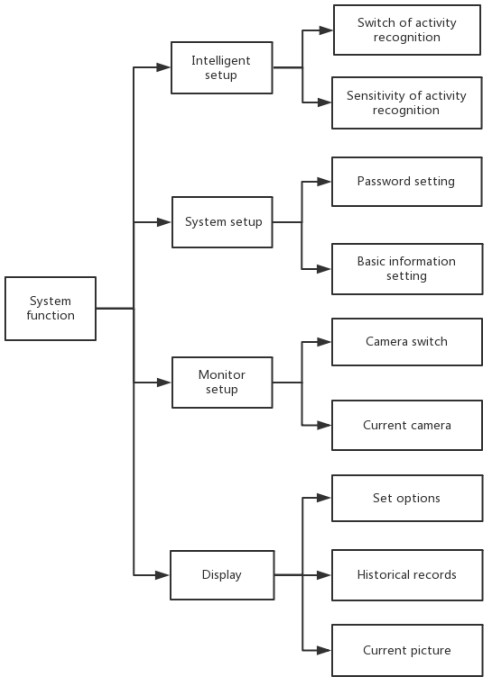

Figure 1. The system function module diagram.

2.1 Monitoring information display

[a] Corresponding author: wangmantao@sicau.edu.cn

All kinds of information collected by the system and the current state of the various settings of the system need to be known by the security administrator. The security administrator can see the current picture taken to monitor and shot this module, they also can look over the history of the activity recognition records and the history of the waring records.

2.2 Other settings

In the settings of monitoring, there are select of monitoring cameras, switches of cameras and so on. In the human activity recognition setting, it includes the switch of activity recognition warning, the update of human activity recognition library, the setting of sensitivity of activity identification, etc. The system settings include the setup and management of account information. Such as the unit name, address, login password, and so on.

3 System structure design

The indoor security video surveillance system based on human activity recognition mainly includes four parts: video data acquisition by camera; construction of human activity recognition and analysis engine; processing and storage of video data by server; design of front-end page. The core as well as difficulty of the system lies in the construction of the human activity analysis engine. The four parts of the whole system are independent but still closely related. The first step of the system is to use a camera in the data acquisition section and collect a large number of experimental personnel to simulate the video fist, kicking, landing, waving, jogging, walking, talking and jumping. And then combining the public data set KTH, each video clip is placed in the folder by category. In the part of the human activity recognition engine, the training set is put into the network for training. When the model training is completed, the generated H5 file can be put into the program on the server. On the server side, to write the page configuration file, the database configuration file, and write the Java file according to the Spring MVC framework. The video stream is introduced into the server in real time when the system is running, and the video is cached and analysed on the server. The results are stored in the database and read on the web page. The structure of campus safety monitoring system based on human activity recognition is shown in Figure 2.

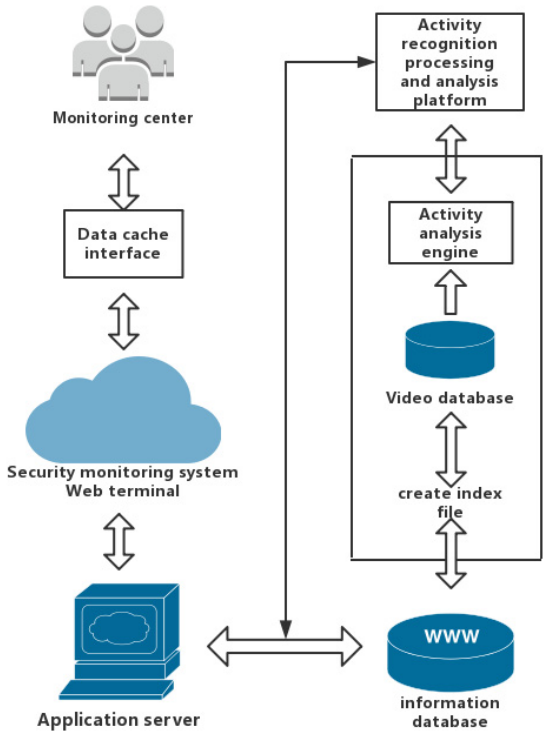

Figure 2. System structure diagram.

3.1 Data acquisition architecture design

The main source of data in the system is the existing video surveillance system. The main purpose of the system is to warn the abnormal human actions, so the real-time requirement is very high. Early most of people use analogue video surveillance system, now the digital video surveillance system is used widely. The transmission of digital video surveillance can be transmitted by wireless or wired [1]. The digital video surveillance system uses embedded video web [2] in the data collection. The video signal which sent by the camera can be directly transmit to the server after the compression of the video signal. In this way, all the devices can be identified by IP address, directly connected to the LAN, without the limit of the length of the cable. It can easier arrange the complex monitoring network in the building, and it has a good expansibility, because the increase of the equipment is only an increase of IP. The architecture of data acquisition is shown in Figure 3.

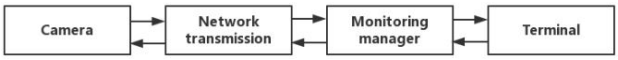

Figure 3. Data collection structure diagram.

3.2 Human activity recognition engine architecture design

This system needs to analyse the video stream and use the idea of image recognition. At present, the mainstream human activity recognition methods are based on local feature representation and deep neural

network [3]. The former is subdivided into interest point detection, local feature extraction and local feature aggregation. The latter includes space-time network, multi stream network [4], depth generation network and temporal consistency network. By comparing the results of different networks in different data sets [5], VGG16 is chosen as the main body of the system neural network.

The comparison of the performance of different networks is shown in Table 1.

Table 1. Test results of different convolution neural networks in main data sets.

Method	VOC-2007 (mean AP)	VOC-2012 (mean AP)	Caltech-101 (mean CR)	Caltech-256 (mean CR)
Zeiler & Fergus (Zeiler & Fergus, 2013)	-	79.0	86.5 ± 0.5	74.2 ± 0.3
Chatfield et al. (Chatfield et al., 2014)	82.4	83.2	88.4 ± 0.6	77.6 ± 0.1
He et al. (He et al., 2014)	82.4	-	93.4 ± 0.5	-
Wei et al. (Wei et al., 2014)	81.5	81.7	-	-
VGG Net-D (16 layers)	89.3	89.0	91.8 ± 1.0	85.0 ± 0.2
VGG Net-E (19 layers)	89.3	89.0	92.3 ± 0.5	85.1 ± 0.3
VGG Net-D & Net-E	89.7	89.3	92.7 ± 0.5	86.2 ± 0.3

VGG16 is a large convolution network. The number of neurons contained in it is shown in Table 2.

Table 2. Parameters contained in each layer of the VGG16 network

conv3-64, conv3-64	3*3*3*64+3*3*64*64+64*2	38720
conv3-128,conv3-128	3*3*64*128+3*3*128*128+128*2	221440
conv3-256,conv3-256,conv3-256	3*3*128*256+3*3*256*256*2+256*3	1475328
conv3-512,conv3-512,conv3-512	3*3*256*512+3*3*512*512*2+512*3	5899776
conv3-512,conv3-512,conv3-512	3*3*512*512*3 + 512*3	7079424
fc-1	512*7*7 * 4096 + 4096	102764544
fc-2	4096*4096 + 4096	16781312
fc-3	4096*1000+1000	4097000
	Total parameters	**138,357,544**

As shown in the table, the final output of the VGG network has more than 138 million parameters. Confront such a large number of parameters, in order to build this system quickly, all layers in the VGG16 network are frozen and loaded into the weight file which does not contain the top weight. In the end, the system is added to the layer of the classification.

3.3 Server architecture design

The system architecture is designed to use MVC framework to separate models, views and control layers, and implement [6] based on Spring MVC framework. Spring MVC is a lightweight Web framework, which easy to use, with a large number of technical documents, little difficulty, and good extensibility. It often used to build a high quality Web application. Tomcat [7] is selected at the bottom of the Java EE Servlet server. The system has both user and video database, user database stores users' username, password and other related information. What's more about video database, the system doesn't store video files into the storage of the server, only the path of video files is saved in the database. Finally, the database configuration file is configured on the server to host the connection of the database to Spring. The basic flow of the whole service program is shown in Figure 4.

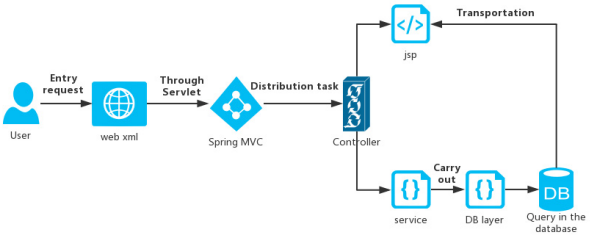

Figure 4. Basic flowchart of service program running.

4 System implementation

The modules implemented on the embedded video server including video scheduling and transmission module, real-time acquisition module, control information processing module as such. The design of data acquisition part is shown in Figure 5.

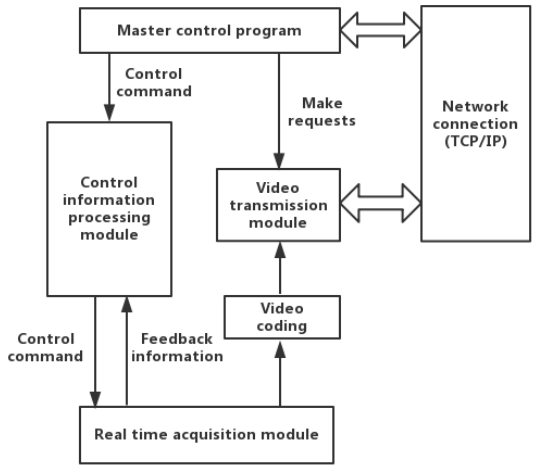

Figure 5. Hardware design of data acquisition part.

The video transmission module is the most important part of the data collection. It selects the corresponding scheduling strategy according to the service type to create the video stream. The video data is packaged and sent to the server. The transmission of video data is selected as the UDP protocol. UDP is often used to transmit data in the monitoring system, and the use of UDP can greatly guarantee the QoS. The multicast of UDP protocol can send packets to a specific user group. This system adopts IP multicast technology.

4.1 Build an engine for activity recognition

The main technologies used in this system are 3D convolution, migration learning, VGG-16 network and data enhancement. Because of the data set in hand is limited, we build it with the public data set KTH, and use migration learning and data enhancement technology to help train a better network.

4.1.1 3D convolution

When the image is input to the original image, images are not processed into a grayscale image. The grayscale image level is two-dimensional, but the RGB image is a stack of three-color layers of red, green and blue, which has three-dimensional blocks. So we can't use 2D CNN here, we construct convolution neural network which convolution the three dimensions by using the idea of 3D CNN [9]. The original image can be seen as a large rectangle with three slices reclosing. The filter is a small cube which is reclosing with three slices. The number of two object channels in the three-dimensional convolution are same. At this time, the small cube can be put into the large rectangular body to be calculated.

The height and width of the picture in system are both 224 pixels, the height and width of the filter are both 3 pixels, and each layer is a two-dimensional matrix. It is assumed that the three layers of the image are R, G and B respectively. The elements of a matrix are: $R_{i,j}(i=1,2,\ldots,224 \quad j=1,2,\ldots,224)$, $G_{i,j}(i=1,2,\ldots,224 \; j=1,2,\ldots,224)$, $B_{i,j}(i=1,2,\ldots,224 \; j=1,2,\ldots,224)$. The three layers of filter correspond to R, G and B respectively is r, g and b. The elements of the filter matrix are: $r_{i,j}(i=1,2,3 \quad j=1,2,3)$, $g_{i,j}(i=1,2,3 \quad j=1,2,3)$, $b_{i,j}(i=1,2,3 \; j=1,2,3)$. Assuming that the output matrix is $S_{i,j}$, the value of the element $S_{1,1}$ of the output matrix can be calculated as (1):

$$S_{1,1} = \sum_{i=1}^{3}\left(\sum_{j=1}^{3}\left(R_{i,j} \times r_{i,j} + G_{i,j} \times g_{i,j} + B_{i,j} \times b_{i,j}\right)\right) \quad (1)$$

Assuming that the step length is 1 pixel, one grid is moved each time, and the above calculation is carried out after moving. Finally, a 222 * 222 size two-dimensional matrix will be obtained. A filter can only extract one feature in the picture, and use multiple filters to convolution separately, so that multiple two-dimensional images can be output. By stacking these two-dimensional images, a new multidimensional image will be generated. The number of channels used in this image equals the number of filters used.

4.1.2 Migration learning

This system uses the open source VGG16 convolution neural network [5] for human activity identification, which uses the idea of three-dimensional convolution when building the network. The number of channels in the image is three, which corresponds to three colours of red, green and blue respectively. The network is built with 16 layers, and the input original image size is 224 * 224 * 3, and at the beginning there are two coiling layers that use 64 * 3 * 3 and a stride with a step of 1. Next, the picture is compressed with a pool layer of 2, the output size is 112 * 112 * 64, and then a number of coiling layers and a pool layer are intersecting. After 5 rounds, the feature graph is fully connected and finally activated by Softmax. The network structure of VGG16 is shown in Table 3.

Table 3. VGG16 network structure

VGG-16 Net					
A	LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input					
conv3-64	conv3-64 LRN	conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
Maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
Maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
Maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
Maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
Maxpool					
Fully connected-4096					
Fully connected-4096					
Fully connected-1000					
Softmax					

The system use PyCharm to build the convolution neural network, and the process of building VGG16 is as follows. Keras package is convenient in PyCharm to build the network structure. Keras is an open source deep learning library based on python. This system uses Keras to help building convolution neural network.

The method of migration learning in this system is freezing all layers in base-model, train and adjust the weight only by the newly added layer. The method of adding new layers and freezing all parameters in the VGG16 network in this system is:

```
def add_new_last_layer(base_model,nb_classes):
    x=base_model.output
    x=GlobalMaxPooling2D(name='pooling')(x)
    x=Dense(fc_size,activation='relu')(x)
    predictions=Dense(nb_classes,activation='softmax')(x)
    model=Model(input=base_model.input,outputs=predictions)
    return model

def setup_to_transfer_learn(model,base_model):
    for layer in base_model.layers:
        layer.trainable=False
    model.compile(optimizer='rmsprop',
                  loss='categorical_crossentropy',
                  metrics=['accuracy'])
```

There are ten types of human actions that need to be identified in this system: hugging, jogging, walking, talking, jumping, punching, kicking, falling down, waving, asking for help and pushing. In recognition, the input video is divided into frames, and the OpenCV library function cv2.imwrite is called in Python to save the video stream by frame. The location of the picture is set to the location where the video frame is saved.

4.2 Build the server side of the system

In this system, the processing of video streams and the building of system pages are all running on the server. In order to deploy the system more quickly, system using Browser/Server structure [10], all the operations that campus security administrators need to accomplish are all implemented through browsers. The development tool used in this system is Intellij IDEA 14.0.3. Using Intellij to set up Javaweb projects, CSS, JS, page and data related directories in the folder under the project. In the folder, a WEB-INF folder and the web.xml file is set up, then the Spring MVC framework controller is introduced and configured. Finally, the JSP file is written under the directory Web Content for the service page, which includes the login.jsp file, index.jsp file, view.jsp and other file application requests.

In addition, the system uses Spring-security to support system security [11], and provides security services for Web application through Servlet filter.

4.3 System page design

This part uses HBuilder for page design. HBuilder provides code input, programmable code block, large Grammar Library and compatible library for syntax browsers. It can meet the needs of monitoring human action identification system applications. Besides, using HBuilder design page has a very great user experience, greatly improving the efficiency of developer.

5 Concluding remarks

The completion of human activity recognition in video surveillance system is an important step to realize the intellectualization of video surveillance. In this paper, the convolution neural network is used to identify the human action, and the structure of the activity recognition video monitoring system is designed. It can use its own data set, specify the behaviour types that expected recognition. The workload of the training also reduced because of combining with the migration learning. Through the design of the system, the implementation scheme of indoor security monitoring system based on human activity recognition is provided.

Acknowledgments

This work was supported in part by the Youth Fund of the Sichuan Provincial Education Department under Grant 18ZB0467 and in part by Research interest training program of Sichuan Agricultural University in 2016 under Grant 04054593.

References

1. Malhi, Karandeep, et al. "A Zigbee-Based Wearable Physiological Parameters Monitoring System." *IEEE Sensors Journal* 12.3(2012):423-430.
2. Alvarezcampana, M., et al. "Smart CEI Moncloa: An IoT-based Platform for People Flow and Environmental Monitoring on a Smart University Campus. " *Sensors* 17.12(2017):2856.
3. Herath, Samitha, M. Harandi, and F. Porikli. "Going Deeper into Action Recognition: A Survey *, **." *Image & Vision Computing* 60(2017)
4. Simonyan, Karen, and A. Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos." *Computational Linguistics* 1.4(2014):568-576.
5. Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *Computer Science* (2014).
6. Ladd, Seth, and B. Smeets. *Building Spring 2 Enterprise Applications*. Apress, 2007.
7. Brittain, Jason, and I. F. Darwin. *Tomcat: The Definitive Guide*. O'Reilly Media, Inc. 2003.
8. Zi-Jing, X. U., and E. S. Division. "The Design and Development of the Network Monitoring Software Based on UDP." *Computer Knowledge & Technology* (2014)
9. Xu, Wei, et al. "3D Convolutional Neural Networks for Human Action Recognition." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 35.1(2012):221-231.
10. Wu, Canglin. "The Analsis and Comparison Between Browser/Server Structure and Client/Server Structure." *Computer Study* (1999).
11. Deinum, Marten, et al. *Pro Spring MVC: With Web Flow*. Apress, 2012.