Research on Image Processing Technology of Dynamic Face Video Recognition Based on Computer Vision Monitoring

Min Liu
Guangxi Vocational & Technical Institute of Industry, Nanning, Guangxi 53001, China
8159006@qq.com

Abstract. At present, face recognition methods are mainly used to identify static images. In surveillance video, faces in different video frames are correlated and only part of the face can effectively reflect face information; according to the characteristics of face image changes in surveillance videos, a face recognition method based on surveillance video is proposed; firstly, a video face sequence is obtained by combining face detection and tracking technology, and then a partial face image recognition result in the video face sequence is taken as a guide to select all face sequence images. The representative face image in the image is recognized, and finally the face information is comprehensively reflected according to the recognition results of all the selected face images; the experimental results show that this method can effectively improve the face in the surveillance video while ensuring the recognition rate and the false recognition rate. Real-time recognition.

Keywords: video surveillance, dynamic face video recognition, image processing technology, software design, computer vision

1. Introduction
In order to effectively monitor and pre-warning to prevent the social harm caused by terrorist attacks and various criminal activities, and to protect the safety of the country and people's lives and property, security agencies in various countries have widely applied video surveillance technology to various fields and places of security prevention. Make it an important part of the anti-terrorism, explosion-proof, anti-theft, and anti-emergency security system. Video surveillance video is not only an important means of public security and crime prevention, but also a nemesis of all kinds of criminals, providing direct on-site evidence for criminal investigations and judicial trials. However, due to the restrictions and influences of various external conditions, most surveillance images are unclear and unrecognizable and lose their application value. In our previous inspection work practice, we often received crime scene surveillance videos sent by the criminal investigation department for inspection, and more than 80% of them were unable to identify individuals due to the mask or unclear images of the suspect in the video, thus losing the evidence function [1].
2. Face recognition technology

Face recognition technology mainly includes three parts: face detection, face feature extraction, and face recognition and comparison. The common face recognition system is shown in Figure 1. Face detection that is, judging whether there is a face image based on whether there is data that meets the face characteristics, which is the basis of face recognition technology. Face feature extraction is to use a certain method to represent the detected face image data information. The quality of face feature extraction directly affects the overall results of the face recognition system. Face recognition and comparison is to calculate and match the detected face with the known face data in the image database to obtain relevant information such as the person's identity. The core lies in the selection of an appropriate face feature representation method to match the face. Therefore, the method of face comparison and recognition is inseparable from the expression of face features [2].

![Diagram of Face Recognition System](image)

Figure 1. Face recognition system

2.1. Face detection method based on geometric features

The earliest face detection methods include methods based on geometric features. In this method, the detected meaningful face features and the distance between them, as well as other geometric features are combined and stored in a feature vector to represent the face. The geometric features of a human face mainly refer to the distribution of various organs of the human face. For example, in a frontal human face, the eyes are always symmetrical, and the nose and mouth are always vertically symmetrical. At this time, face detection can be performed based on these rules describing the distribution of facial features, that is, prior knowledge of the topological and geometric relationships of the face. As long as the feature vector of the image to be tested is compared with the feature vector of the known face obtained in the database, the distance between the two vectors can be obtained by the similarity measurement to determine whether the image to be tested is a face image [3].

2.2. Face detection method based on skin color model

The skin color of a human face can be better concentrated in a small area of the color space, so it can be described with the help of color space information, such as color space: RGB, HSV, YCrCb, etc. Although different people have different skin colors, studies have shown that the main difference in skin color lies in its brightness, not its chromaticity. Therefore, face detection can be performed with the information obtained by the skin color model. For example, the Gaussian model based on YCrCb color space can effectively separate brightness and chroma, and obtain good clustering. The following two-dimensional Gaussian functions are commonly used to express skin color distribution:

\[ P(Cb, Cr) = \exp\left[-0.5(x-M)^T C^{-1}(x-M)\right] \]  

(1)
However, since other exposed limbs of the human body will also interfere with the face detection result based on skin color, such as the arm of the human body, the face detection method based on the skin color model is usually combined with other face detection methods. That is, the skin color model is used to detect and lock the area where the face may be located, and then for this area, further more accurate face detection is performed.

3. The overall framework of video face recognition

The overall recognition structure of the face recognition method based on surveillance video is mainly composed of two parts, namely, the video face sequence image acquisition part and the video face sequence image recognition part. As shown in Figure 2, the left part obtains the sequence of face images in the video stream by combining face detection and tracking and stores them in the face sequence buffer space, and the right part uses the query buffer space to coordinate the complete face sequence in the buffer space recognition.

![Figure 2. The overall framework of video face recognition](image)

In the video face sequence image acquisition part, the face position and size information in the video frame is obtained through face detection. According to the detection results, the subsequent video frames are tracked. After N frames are tracked, the detection is performed again and the tracking face information is corrected and updated according to the detection results. In this way, a sequence of face images is alternately obtained. The video face sequence image recognition part, by constantly querying whether there is a complete face sequence image generation, the complete face sequence image uses the time correlation between the face sequence images to perform collaborative recognition to obtain the recognition result of the corresponding face sequence [4].

3.1. Selection of face detection and tracking algorithms

The realization object of video face sequence collaborative recognition is the complete face sequence of the target face in the video, and the function of face detection and face tracking is to provide an effective face sequence for it. This paper combines the existing face detection algorithm and face tracking algorithm to obtain the complete face sequence of the target face in the surveillance video. The face detection algorithm uses K. P. The MTCNN algorithm proposed by Zhang et al. achieves a better face detection effect through a well-designed cascade CNN architecture, online difficult sample mining, and face alignment joint learning. The face tracking algorithm uses the Correlation Tracker single-target tracking algorithm in Dlib. This algorithm has a good balance between accuracy and real-time performance by using a method used in a detection frame to perform robust scale estimation. It can effectively track human faces in general scenes.
3.2. Face sequence generation

The purpose of tracking face detection and tracking is to obtain all face sequence images of each face appearing in the surveillance video. The main realization idea is that the video is detected every N frame, and the face information obtained from the detection is tracked between adjacent detection frames, and the tracking results are corrected for each detection until the face disappears to obtain a complete face sequence image. The specific steps are as follows:

Step 1: The initial detection obtains all face position information in the current frame, and starts tracking with the face detection result as the tracking target and stores the corresponding face sequence image in the face sequence buffer space.

Step 2: After tracking N frames, detect the video frame again to obtain the detected face information. Set the area of the detected face area as D, and track the face area Ti (i represents the ith tracked face), and the area of the detected face area and the tracked face area overlapped is. Let P be the maximum coincidence degree between the detected face and D ∩ Ti, all tracked faces, and P is defined as follows:

\[ P = \max \left[ \frac{D \cap T_i}{\min \{D, T_i\}} \right] \]  

(2)

If P is greater than the set threshold, it is considered that the current detected face belongs to P and corresponds to the tracking face i. When tracking in the next frame, use the detected face information instead of tracking the position information of the face for tracking. If P is less than the set threshold, increase tracking of faces.

Step 3: Use Step2 to correct the tracked face or increase the tracked face for all the detected face information.

Step 4: Repeat the above steps to detect and track until the face disappears in the video, and add an end mark to the face sequence that ended tracking.

4. Facial feature value extraction

4.1. Formation of the overall dispersion matrix

For the H×W N face image set \( \{F_1, F_2, \cdots, F_N\} \), according to the specific conditions of the image during the imaging process, it is usually necessary to perform certain preprocessing first, which may include image processing steps such as image smoothing and grayscale enhancement. Then, a necessary process is to expand the image in matrix form into a \( HW \times N \) -dimensional vector in rows or columns, namely

\[ \{F_1, F_2, \cdots, F_N\} \Rightarrow \{x_1, x_2, \cdots, x_N\} \]  

(3)

In order to make its mean value zero, you need to subtract its mean vector, that is

\[ X = \{x_1 - \mu, x_2 - \mu, \cdots, x_N - \mu\} \]  

(4)

Where X is called the training sample matrix. \( \mu \) is the estimate of the statistical mean of X, namely

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]  

(5)

At this time, the autocorrelation matrix R in the K-L transformation becomes the covariance matrix \( S_x \), and the maximum likelihood estimation can be obtained as follows (ignoring the constant 1/N)
This estimation is also called the overall dispersion matrix, which is the generator matrix of the K-L transformation.

4.2. Dimension problem
When the H×W face image is represented as a vector, its dimension is \( n = HW \), and thus the dimension of the overall dispersion matrix is \( n \times n \). For example, if \( H = W = 100 \), then the overall scatter matrix \( S_T \) is dimensional, and the number of samples in the face database \( N \) is usually much smaller, for example, several thousand or so, which leads to the rank of the overall scatter matrix is determined by \( N \). It can be seen from equation (6) that in the \( N \) terms of the sum formula, only \( N-1 \) terms are independent. Therefore, the rank of the overall scatter matrix is usually \( N-1 \), which is also the number of its non-zero eigenvalues. The feature vector corresponding to the zero-feature value is meaningless in the representation of the face image. This can be obtained from the discussion of feature compression in Section 4.2. In summary, in the PCA method of human face, the upper limit of the number of selected feature vectors should be \( N-1 \).

4.3. Solution of transformation matrix and realization of transformation
Solving the transformation matrix is to find the eigenvalues and eigenvectors of the generator matrix. According to (5) and (6), there are

\[
S_T \Phi = \Phi D \tag{7}
\]

Or

\[
S_T \varphi_i = \lambda \varphi_i \tag{8}
\]

Generally, there are three ways to solve the characteristic equation:

4.3.1. Direct solution method. Solve directly the algebraic equation in (7) or (8). Since the dimensionality of the overall scatter matrix is usually very high, this solution method is time-consuming. The following functions are available in MATLAB

\[
[u, v] = \text{eig}(S_T) \tag{9}
\]

The above MATLAB function does not sort the obtained eigenvalues.

4.3.2. One of the indirect solutions. Different from the direct solution method, here is not to directly find the eigenvalues and eigenvectors of \( S_T \), but to obtain the eigenvalues and eigenvectors of \( X^T X \). Specifically, to solve the eigenvalues and eigenvectors of \( X^T X \), there are the following characteristic equations

\[
X^T \Phi \Phi^T \Phi = \Phi \Phi^T \Phi D \tag{10}
\]

After multiplying both sides of the above formula by \( X \), there is

\[
(XX^T \Phi ) \Phi = \Phi (XX^T \Phi) \tag{11}
\]

Comparison (11) shows that \( \Phi = XX^T \Phi \), \( D_\lambda = D_\lambda \). Obviously, from the discussion in Section 4.2, we can see that the dimension of \( X^T X \) is \( N \times N \), which is usually much lower than the dimension of \( XX^T \).
4.3.3. Indirect solution method two. By performing singular value decomposition (SVD) on the training sample matrix \( X \), the characteristic equation can also be solved indirectly. The SVD of \( X \) can be expressed as

\[
X = PTQ^T \\
\left( X \in R^{m \times N}, P \in R^{m \times m}, T \in R^{m \times m}, Q \in R^{N \times N} \right)
\]  

(12)

Where \( P \) and \( Q \) are orthogonal matrices, and \( T \) is a singular value matrix, so the overall dispersion matrix can be written as

\[
XX^T = PTQQ^T T^T P^T = PTT^T P^T
\]  

(13)

After further finishing

\[
XX^T P = PTT^T = PA
\]  

(14)

The following functions can be used for singular value decomposition in MATLAB

\[
[u,s,v] = svd(X,0)
\]  

(15)

The above MATLAB function will sort the obtained singular values. Once the eigenvalues and eigenvectors are obtained, the eigenvectors corresponding to the first \( M \) largest non-zero eigenvalues can be selected according to the method of formula (15) to form the transformation matrix \( \Phi \), and KL is performed on all face samples according to the following formula Transform.

\[
y_j = \Phi^T(x - \mu)
\]  

(16)

4.3.4. Eigenface. We sort the feature values from large to small: \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \), its corresponding feature vector is \( \varphi_1 \geq \varphi_2 \geq \ldots \geq \varphi_m \). In this way, each face can be projected into the subspace formed by \( \varphi_1 \geq \varphi_2 \geq \ldots \geq \varphi_m \), so each face image corresponds to the subspace A point in space. Similarly, a point in the subspace also corresponds to an image. When the feature vector \( \varphi_1 \geq \varphi_2 \geq \ldots \geq \varphi_m \) is displayed as images, these images are very similar to human faces, so they are called "eigenfaces". With such a feature subspace formed by "feature faces", any face image can be projected to it and a set of coordinate coefficients can be obtained. This set of coefficients indicates the position of the image in the subspace, thus it can be used as a basis for face recognition. In other words, any face image can be expressed as a linear combination of this group of "eigenfaces", and its weighting coefficient is the expansion coefficient of K-L, which can also be called the algebraic feature of the image. Figure 3 shows the images of 8 people in the selected ORL face database, and Figure 4 shows the first five feature faces.

![Image 1](Image1.png)  

**Figure 3.** Images of 8 people in the ORL face database
5. Design of face recognition system based on video surveillance

5.1. System function goals
Stability is an important goal pursued by this system, so servers and webcams are separated to ensure the stability of the service system. The collection and transportation of video data is the main work of the webcam, and the server is an important hub for connecting working equipment. It is connected with the client, webcam and other equipment, and has the functions of storing information in the distribution control instructions [7-8]. This video surveillance system is mainly composed of client terminals, servers and related camera devices. In order to further improve the portability and security of the network monitoring of this system, the client terminal does not have the authority to directly access the network monitoring camera equipment. First, the server receives the information transmitted by the network monitoring equipment, and modulates and decodes the information of the monitoring equipment. The resulting video stream is finally sent to the terminal device for analysis. In the actual working process, the system mainly uses face recognition technology, and finally connects to the security prevention and control system. The center of the system is the network server. The network server also has the storage, analysis, and monitoring of wireless sensor network data. Store monitoring data and manage the role of identifying and analyzing staff identity information. The overall structure diagram of this system is shown as in Fig. 5.

![Figure 5. The overall structure of the monitoring platform for security monitoring sites](image-url)
5.2. Server business logic design

The structure of the monitoring system is shown in Figure 6. From a logical point of view, the system can be divided into two independent modules, so the video monitoring system uses a dual-process, multi-threaded design mode. A process is mainly for the collection and transmission of video images of surveillance camera equipment. The server side intermittently sends request messages to the network camera device, and receives the data information transmitted from the terminal camera device. It can be seen that the system adopts is the full-duplex transmission mode. The image information collected from the terminal camera equipment is transmitted to the face recognition unit, and the image is compared and analyzed with the picture in the database, and the picture is decoded to form a media data stream in the corresponding format, and then the RTP routing transmission protocol in the computer network Send streaming media data to the client terminal. The main job of another process is the connection and communication of the terminal device. The first step is to initialize and create an object. If a connection request is generated, a thread is called from the thread pool to realize the processing, and the data in the database is updated in real time. When the process ends, the resource is recovered through the recycling mechanism. This system uses a semaphore mechanism to prevent deadlocks, and uses pipes to achieve communication between processes.

![Data flow chart of video surveillance](image)

Figure 6. Data flow chart of video surveillance

5.3. Client business logic design

The terminal uses QT Creator software for development, and uses QT signal and slot mechanism to respond to events. The user directly enters the system from the client interface, and the network monitoring module and wireless sensor network exist in the security monitoring platform.

5.4. System test

5.4.1. Recognition time test. This test determines the time period from the time someone enters to the confirmation of their identity. This judgment time will directly determine the efficiency of the system.
Although low efficiency does not affect accuracy, its security performance cannot be guaranteed. The hardware configuration will also have a certain impact on the time of system detection and judgment. The resource configuration information of the development board used in this research and test is shown in Table 1.

| Server-side test environment | Client test environment |
|-----------------------------|------------------------|
| operating system            | Ubuntu12.04LTS         |
| CPU                         | 620MHZ                 |
| Cache                       | 2MCache                |
| RAM                         | 4GB                    |
| hard disk                   | 2TB                    |
| server                      | Think Server           |
| Network IP                  | 192.168.100            |

From the perspective of the entire detection program, most of the time spent will be spent on motion capture, face recognition and detection, and interactive communication. The process of delivering the result of face recognition to the client is called real-time interactive communication. Therefore, after these time-consuming tests, the results of the system detection and judgment time can also be obtained. Set the reference variables T1, T2, T3, T1 is the length of time it takes for the image of the moving object to be captured, T2 is the time required for the entire detection and identification process, and T3 is the time it takes to push data [5].

5.4.2. Motion capture time. When the motion data cannot be captured, the entire thread is in a blocked state, which does not occupy the CPU. Once the motion data appears, SAD will be obtained based on the API, and the motion frame will be obtained after scanning the data row and column. In this test, a total of 100 motion capture experiments were completed to analyze the T1 situation. The results are shown in Figure 8.

![Figure 7. Motion capture time consumption line chart](image)

According to the above chart, the test time consumed by the three stages of motion capture, face recognition, and information transmission is compared and shown in Table 2.
Table 2. Various time consumption statistics

|                                | Maximum consumption time (ms) | Average consumption time (ms) | Minimum consumption time (ms) |
|--------------------------------|--------------------------------|-------------------------------|-----------------------------|
| Motion capture                 | 40                             | 33                            | twenty-four                 |
| Face recognition               | 6658                           | 3402                          | 1210                        |
| Information transfer           | 101                            | 19                            | 11                          |

6. Conclusion
This article discusses the face recognition method based on surveillance video, based on the face recognition method for the image, combined with the time correlation of the face image in the video, by selecting some of the face sequence images for recognition and analysis, to obtain face identity information. The method proposed in the article is oriented by partial face image recognition results, and selects partial sequence regions with optimal recognition results as a method for recognition of representative face images, which provides a new method for the application of image-based face recognition methods in video Ideas. Through experimental verification, the method described in this paper effectively improves the real-time performance of video face recognition under the premise of ensuring the recognition rate, and improves the misrecognition in video face recognition to a certain extent.

References
[1] Campbell, & Sarah. Baby, where did you get those eyes: ieee pulse talks with mark sagar about the new face of artificial intelligence. IEEE Pulse, vol. 6, pp. 4, June 2015.
[2] Hossain, M. S., & Muhammad, G. Cloud-assisted speech and face recognition framework for health monitoring. Mobile Networks and Applications, vol. 20, pp. 391-399, March 2015.
[3] Mantecon, T., Del-Blanco, C. R., Jaureguizar, F., & Garcia, N. Visual face recognition using bag of dense derivative depth patterns. IEEE Signal Processing Letters, vol. 23, pp. 771-775, June 2016.
[4] Yan, X., Valérie Goffaux, & Rossion, B. Spatial frequency tuning of single-glance familiar face recognition in a dynamic visual stream. Journal of Vision, vol. 19, pp. 92, October 2019.
[5] Elgargni, M., Al-Habaibeh, A., & Lotfi, A. Cutting tool tracking and recognition based on infrared and visual imaging systems using principal component analysis (pca) and discrete wavelet transform (dwt) combined with neural networks. The International Journal of Advanced Manufacturing Technology, vol. 77, pp. 1965-1978, September 2015.