Binocular Vision-Based Human Ranging Algorithm Based on Human Faces Recognition

Xiaolin He, Lin Ma(✉), and Weixiao Meng

School of Electronics and Information Engineering,
Harbin Institute of Technology, Harbin, China
17s005066@stu.hit.edu.cn, {malin,wxmeng}@hit.edu.cn

Abstract. In the field of security, timely and effective identification is very important for safeguarding public safety, national security and information security. Face recognition is an important technology in these areas. The calculation of range plays an important role in protecting safety and tracking suspects. Binocular stereo vision ranging has wide application in non-contact precise measurement and dangerous scenes. In this paper, a binocular range measurement system based on face recognition is proposed. The system can detect and recognize faces and calculate its real time range. It could realize tracking real time faces and calculate its distance from the cameras and locate them. And it suits the feature of special places of high security and preventing the suspicious people from entering and out.

Keywords: Face recognition · Binocular stereo vision ranging

1 Introduction

With the increasing demand for fast and effective identity recognition in society, the security problem is of great urgency [8]. At the same time, the recognition technology based on biometrics has gradually become a hot spot. Face recognition technology is generally accepted by people because of its non-contact and friendly interface [10]. In many ways of perceiving the world, visual information takes up about 80%. Binocular stereo vision has been widely applied to various aspects of production and life [5], especially in dangerous scenes.

Although the intelligent video surveillance is more and more mature [13], the pressure of people's demand for intellectualization of video analysis is also growing. However, the current monitoring system only support the storage functions. It relies on people to review video data. Then missing detection or erroneous inspection are easy to happen. At the same time, with the increase of the number of monitoring terminals, the resource consumption of human analysis is becoming more and more difficult to accept. So many researches have introduced computer vision technology in video surveillance in order to realize the intelligent video surveillance system [6].

S. Han et al. (Eds.): AICON 2019, LNICST 287, pp. 414–424, 2019.
https://doi.org/10.1007/978-3-030-22971-9_36

Face detection is originally derived from face recognition. Compared with eigenface and fisherface [3], the most obvious feature of the new algorithms for face recognition is the application of automatic recognition technology in the field of face recognition such as Adaboost [4], support vector machine (SVM) [12]. In the early stage of face recognition research, face recognition needs people to face to the camera. So it does not have to take the background information into consideration such as the location of the face. In this system, due to the long distance from people to camera, it's necessary to detect faces first.

Besides, artificial neural network has excellent performance in machine vision, voice recognition, natural language processing. Convolutional Neural Network (CNN) can classify large image data set very well. In the famous Alphago, CNN is applied to analyze the competition and offer decision information [2].

Computer vision technology makes the monitoring system not only see what happens but also understand what happens [9]. Through improving the analysis technology of video data, the system could recognize and identity the unusual things or people and give the alarm in the fastest and best way [7]. In the system we propose, to realize the intelligence of video monitoring system, the monitoring system can not only provide video data passively, but also analyze and process the video contents automatically.

The main contributions of this article are as follows:

(1) We propose a highly robust system to achieve the recognition and positioning of human faces, and it is real-time.
(2) This system can be applied not only in the general environment but also in dangerous scenes. The experimental results show that this system can effectively identify and locate human faces and meet the accuracy requirements.

The remainder of this paper is organized as follows. In Sect. 2, we will introduce the system model. The algorithms will be discussed in Sect. 3, which conclude face detection and recognition and ranging algorithms. Section 4 will provide the implementation and performance analysis. And conclusion will be described finally.

2 Proposed System Model

2.1 Measurement System Overview

The system contains two parts as shown in Fig. 1: one is offline phase to train the face detector and recognizer. Through the training of a large number of offline data and the establishment of a resource bank, a powerful classifier detector can be obtained, which can save a lot of time and improve efficiency when it is used online; the other is online phase to use them and calculate the real time distance. The system has good adaptability and can be applied to various special occasions.

In the offline, face detector and recognizer are built to detect and recognize faces. AdaBoost algorithm is applied to detect faces and its real time performance and accuracy could satisfy our system. At the same time, Convolutional Neural Network trains the faces and gets the face recognizer.

In the online, binocular vision platform works to capture images and the left image is sent to be tested for face detection. If exists, the corresponding right image is also detected. In the online phase, we apply Speeded Up Robust Features Algorithm to match feature points for distance calculation.

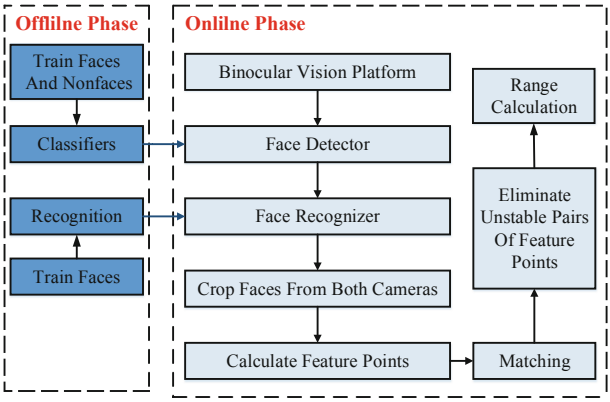

Fig. 1. The flow chart of the system.

2.2 Offline Phase and Online Phase

In the offline phase, the weak face classifier could be obtained by training the faces and non-faces data sets. The results of these weak classifiers can be only slightly better than the results of random guessing. But by cascading these weak classifiers, a strong classifier is obtained.

In the online phase, the binocular vision system will take a series of video streams to the background for processing. After receiving the video stream, the background will firstly detect the face. If there exists human faces, the two faces will be cropped and saved for further process. Then the feature points of the two faces are calculated. The feature points are used to match to calculate the ranging of the faces. The system again removes unstable pairs of feature points before calculating the range of the face. If the person is criminal or someone dangerous, the system will alarm for help to the monitoring platform.

3 Face Recognition and Ranging Algorithm

3.1 Face Detection Algorithm

The first step is to detect faces in pictures. Several methods such as skin color based method are widely used in face detection. AdaBoost algorithm is a main method for face detection and has great advantages over other methods.

It contains thousands of feature matrixes in one picture and the introduction of integral graph can improve the speed and accuracy of detection [11]. Assuming

that a picture is $M \times N$ and the gray value of point (x, y) is $P(\alpha, \beta)$, therefore the integral value is:

$$I(x, y) = \sum_{i=0}^{x} \sum_{j=0}^{y} P_{xy}(\alpha, \beta) \tag{1}$$

Adaboost is applied to combine the rectangular feature after calculating the Haar-like feature. As long as its accuracy is better than the random selected feature, it will be chosen as a weak classifier. Every feature can be trained to be a less better weak classifier. Then through training these weak classifiers, strong classifiers could improve the accuracy. The final classifier is to cascade several strong classifiers.

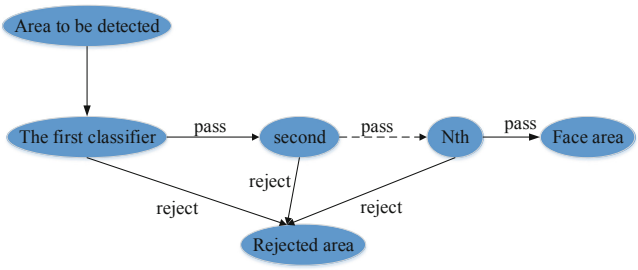

Fig. 2. Cascade classifier.

The cascade algorithm is as shown in Fig. 2. In every classifier, all the regions to be sorting are classified. The one that can reach the threshold will be sent to next classifier. Otherwise, it is considered to be invalid area, and the final face area is obtained after N classifiers. This cascade algorithm can achieve real-time detection effect. Its operation speed is very fast. This is because the face region and the non-face region is extremely asymmetric and the number of areas that can pass the classifier decreases rapidly. So it simplifies the calculation and is very efficient.

3.2 Face Recognition Algorithm

Artificial neural network is designed to imitate the structure of neurons in biological system. It imitates nonlinear processing of information by imitating synaptic connections, structures and functions of brains. The advantage of artificial neural network lies in its high parallel processing, high robustness, high fault tolerance and the correct classification and processing of fuzzy and imprecise information [14]. Convolutional Neural Network develops from multilayer perceptron. It is inspired from the field of biological neuroscience. It simulates the receptive field of the cats' visual cortex. This receptive field is very sensitive to local perception, and the receptive field is tiled to the whole area.

The algorithm flow chart contains 5 main parts:

(1) Input layer. Input layer is the first step of the whole network, and the original image is input directly without so many image preprocessing like other algorithms.
(2) Convolution layer. The convolution layer is the output of the upper layer and the convolution is calculated by a number of convolution cores. Each convolution kernel repeats itself in the entire input region, and the convolution result is a feature graph that forms the input image. The convolution layer is the core of the whole convolution neural network.

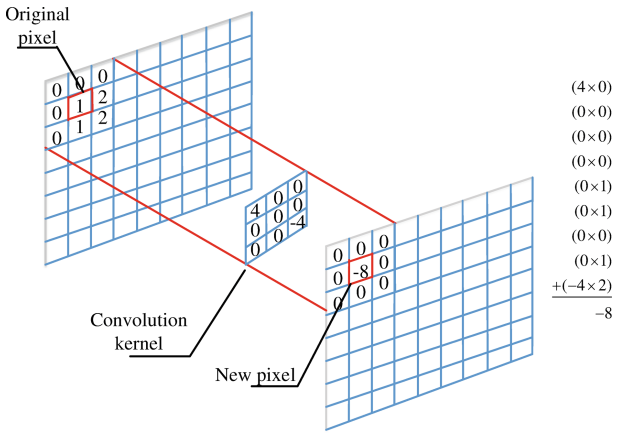

Fig. 3. Convolutional kernel.

Generally a convolution layer contains many convolution kernels (assuming to be n). Every convolution kernel convolutes with all the regions of the picture. Then n corresponding output characteristic maps are obtained. They would be input to the next layer. As shown in Fig. 3, after convoluting the upper image, image data is greatly reduced. However, to keep the size of the same, 0 can be put in the picture.

(3) Pooling. The image scale after convolution layer does not reduce too much. Therefore, pooling is introduced to reduce the dimension of the picture and the computational complexity and enhance the robustness of the network.
(4) Full connection. Many full connection layers are connected to the output layer. And the multiple full connection layers form a shallow layer of multilayer perceptron.
(5) Output layer of softmax. Softmax is the generalization of the logistic regression model and calculates the maximum likelihood probability. It is a common algorithm used to solve multi-classification problem.

3.3 Ranging Method Based on Feature Extraction

The Speeded Up Robust Features Algorithm is based on iteratively detection and extraction of robust local feature points. Compared with other algorithms, the SURF algorithm has a great improvement in speed and robustness, so it is very suitable for real-time stereo matching [1,7]. The SURF algorithm relies on the Hessian matrix when extracting the feature points. More precisely, it depends on the maximum value of the Hessian matrix in the region. Assuming that somepoint X in the picture, its Hessian matrix is defined as:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma)\ L_{xy}(x,\sigma) \\ L_{xx}(x,\sigma)\ L_{xy}(x,\sigma) \end{bmatrix} \tag{2}$$

$L_{xx}(x,\sigma)$ indicates the convolution of the second order derivative of Gaussian $\frac{\partial^2 g(\sigma)}{\partial x^2}$ with the image. So do $L_{xy}(x,\sigma)$ and $L_{yy}(x,\sigma)$. The mathematical expression of $g(\sigma)$ is:

$$g(x,y,\sigma) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma}} \tag{3}$$

The Gauss function needs to be discretized and cut in the actual application. It will be more suitable for the analysis of scale space and reduce the repeat degree of Hessian matrix.

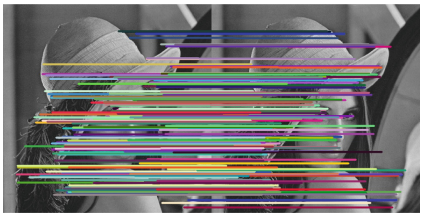

Fig. 4. Match result.

Figure 4 is the result of matching feature points. By using these matched feature points we could calculate its real-time range. The principle of binocular ranging is similar to human eye ranging. So long as the coordinates of the matched feature points are obtained, its range could be calculated. As the Fig. 5 shows, the two cameras are placed in parallel and the horizontal distance is b. One point that should be mentioned is that the two cameras are in the same height. The focal length of the two cameras is f. P is the point in the world coordinate and P_l is the mapped point in the left image and P_r is the mapped point in the right image. (x_l, y_l) is the coordinates of P_l, (x_r, y_r) is the coordinates of P_r. The range could be calculated by the formula:

$$Z = \frac{bf}{x_l - x_r} \tag{4}$$

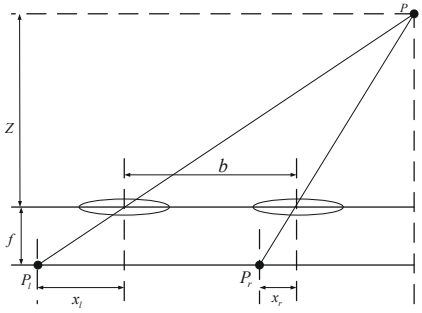

Fig. 5. Binocular ranging principle.

4 Implementation and Performance Analysis

4.1 Offline Training Results

In order to test our proposed method, we make an experiment in our lab, which is located in the Information Building, Science Park of Harbin Institute of Technology, China.

Firstly, offline training is necessary to build face detector and recognizer. To build face detector, 2860 pictures of faces and 4572 pictures of non-faces are used to form the strong classifiers. As shown in Fig. 6, with the increase of the number of weak classifiers, the accuracy rate is increasing. However, with the increase of the weak classifiers, the detection rate will also decrease and the corresponding detection time will increase. So the proper number of classifiers should be selected to balance them.

Secondly, face recognizer is also essential for the system. We use 200 people of their faces pictures and each has 7 pictures of different angles or lights. In the process of training, the number of convolutional kernels in the first layer of CNN is set to be 50, and the second layer to be 70. As shown in Fig. 7 with the increase of learning times, the error rate decreases and finally converges to 3%.

4.2 Online Results

In this experiment, the frequency of the shooting is 5 pairs of pictures per second. Then the pictures of the left and right cameras are sent to detect and recognize. Take the Fig. 8 for example, after human face detection, the position of the face is known. In the left image, (663,417) and (768,536) are the coordinates of the top left corner and lower right corner of the square face area. And in order to ensure real time, only the left image is detected for faces. If there exists, the right image captures the face by default.

Due to the small proportion of the face in the whole image, the size of the face frame is properly enlarged and some part of the body is framed too. It can improve the precision of the feature points matching and avoid the large distance error caused by the lack of matching points.

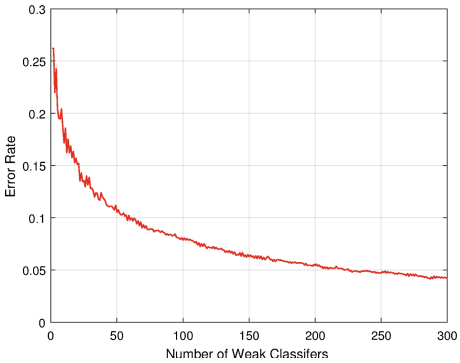

Fig. 6. The error rate of face detector.

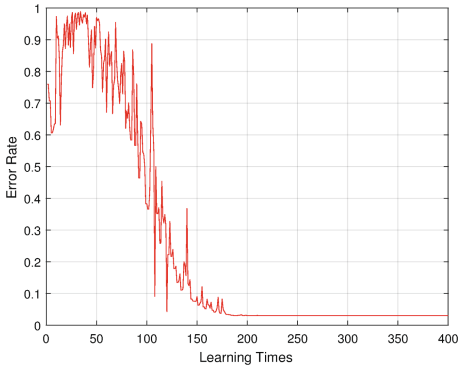

Fig. 7. The error rate of face recognition.

Figure 8 is the final matching result. There are 7 feature points matching pairs.The average distance is 3.64m. While the actual distance is 3.7 m, the absolute error is 0.058 m, and the relative error is 1.6%. The accuracy of the experiment could satisfy the requirement.

Figure 9 is a comparison of the SURF and SIFT algorithms used in the feature point extraction and matching process. From the Fig. 9, the ranging results can be obtained intuitively. Due to the effects of illumination, the SURF algorithm is more suitable for the system than the SIFT algorithm and could match more feature points. And the measurement accuracy is higher. This conclusion can also be drawn from the cumulative distribution function of Fig. 10, and the measurement result of 90% of the measurement results of this system is less than 0.6m, and the relative error of 50% points is less than 10%.

Fig. 8. Matching result.

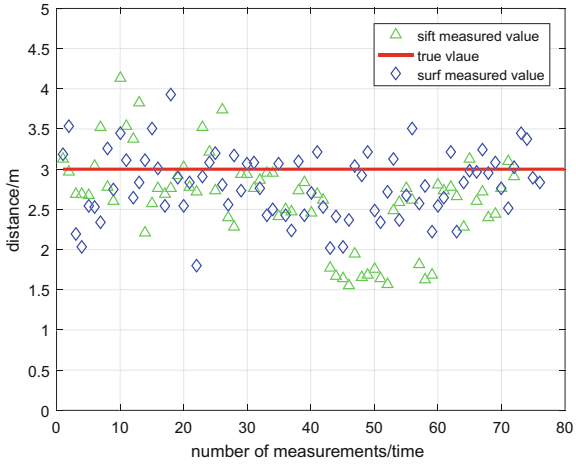

Fig. 9. Measured range.

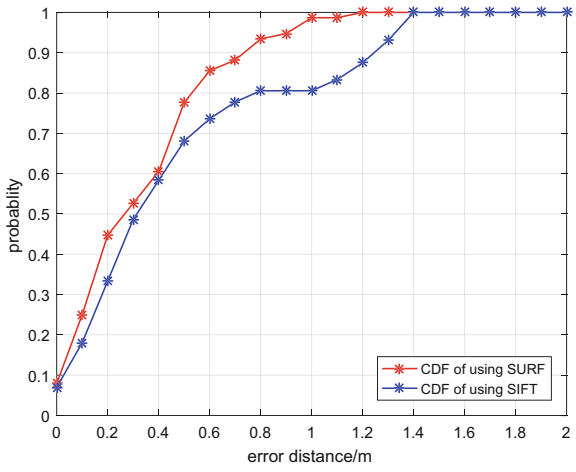

Fig. 10. CDF.

5 Conclusion

In this paper, a binocular distance measurement system based on face recognition is proposed. The AdaBoost algorithm classifies faces and non-faces. CNN is used to recognize different people. SURF algorithm can pick out feature points and match them. Therefore the system could realize tracking real time faces and calculate its distance from the camera and locate them. And it suits the feature of special places of high security and prevents the suspicious people from entering and out to start early-warning.

Acknowledgment. This paper is supported by National Natural Science Foundation of China (61571162, 4181101180), Ministry of Education - China Mobile Research Foundation (MCM20170106) and Heilongjiang Province Natural Science Foundation (F2016019).

References

1. Vinay, A., Hebbar, D., Shekhar, V.S., Murthy, K.N.B., Natarajan, S.: Two novel detector-descriptor based approaches for face recognition using sift and surf. Proc. Comput. Sci. **70**, 185–197 (2015). https://doi.org/10.1016/j.procs.2015.10.070
2. Aloysius, N., Geetha, M.: A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP), pp. 0588–0592, April 2017. https://doi.org/10.1109/ICCSP.2017.8286426
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997). https://doi.org/10.1109/34.598228
4. Burduk, R.: The adaboost algorithm with linear modification of the weights. In: Choraś, M., Choraś, R.S. (eds.) Image Processing and Communications Challenges 9, vol. 681, pp. 82–87. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-68720-9_11
5. Chaisorn, L., Wong, Y.: Video analytics for surveillance camera networks. In: 2013 19th IEEE International Conference on Networks (ICON), pp. 1–6, December 2013. https://doi.org/10.1109/ICON.2013.6782002
6. Gao, C., Li, P., Zhang, Y., Liu, J., Wang, L.: People counting based on head detection combining adaboost and CNN in crowded surveillance environment. Neurocomputing **208**, 108–116 (2016). https://doi.org/10.1016/j.neucom.2016.01.097. sI: BridgingSemantic
7. Kokila, R., Sannidhan, M.S., Bhandary, A.: A novel approach for matching composite sketches to mugshot photos using the fusion of SIFT and SURF feature descriptor. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1458–1464, September 2017. https://doi.org/10.1109/ICACCI.2017.8126046
8. Panagiotou, N., et al.: Intelligent urban data monitoring for smart cities. In: Berendt, B., et al. (eds.) ECML PKDD 2016. LNCS (LNAI), vol. 9853, pp. 177–192. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46131-1_23
9. Ranganatha, S., Gowramma, Y.P.: An integrated robust approach for fast face tracking in noisy real-world videos with visual constraints. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 772–776, September 2017. https://doi.org/10.1109/ICACCI.2017.8125935

10. Tsai, H.C., Wang, W.C., Wang, J.C., Wang, J.F.: Long distance person identification using height measurement and face recognition. In: TENCON 2009–2009 IEEE Region 10 Conference, pp. 1–4, January 2009. https://doi.org/10.1109/TENCON.2009.5396069
11. Wagh, K.V., Kanade, S.S.: Pedestrian detection using integral channel detection and ADABOOST algorithm, Waknaghat, Shimla, India, pp. 383–387, January 2018
12. Xu, J., Zeng, W., Lan, Y., Guo, J., Cheng, X.: Modeling the parameter interactions in ranking SVM with low-rank approximation. IEEE Trans. Knowl. Data Eng. 1 (2018). https://doi.org/10.1109/TKDE.2018.2851257
13. Yu, X., Ganz, A.: Mass casualty incident surveillance and monitoring using identity aware video analytics. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 3755–3758, August 2010. https://doi.org/10.1109/IEMBS.2010.5627536
14. Zarandy, A., Rekeczky, C., Szolgay, P., Chua, L.O.: Overview of CNN research: 25 years history and the current trends, Lisbon, Portugal, pp. 401–404, July 2015