A Hybrid Framework for Human Face Detection and Recognition in Videos

Xueying Huang*
School of Automation, Wuhan University of Technology, Wuhan 430070, China

*Corresponding author e-mail: xueying_huang@yeah.net

Abstract. All manuscripts must be in English, also the table and figure texts, otherwise we cannot publish your paper. Please keep a second copy of your manuscript in your office. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. Should authors use tables or figures from other Publications, they must ask the corresponding publishers to grant them the right to publish this material in their paper.

1. Introduction

With the wide use of cameras in surveillance and mobile devices, a large number of videos are constantly captured. Compared with static images, videos often contain more information, such as time and different perspectives. Video surveillance in public places provides a powerful guarantee for social security and law enforcement. Therefore, the intellectualization of video surveillance system is one of the important research directions in the field of computer vision, especially the application of face recognition technology in surveillance video. [1].

The steps of face recognition mainly include four parts: face detection, face segmentation, face feature extraction and face matching. At present, face recognition algorithms are mainly divided into the following categories: feature-based method, template matching method, appearance-based method and deep learning based method. The method mainly uses manual feature extraction for face recognition, such as Histogram of Gradient (HOG) [2], Local Binary Pattern (LBP) [3], Scale Invariant Feature Transform (SIFT) [4], Gabor [5] and so on. The method based on template matching is to build a standard face model, define the face as a function [6]. In recent years, face recognition methods based on deep learning have become popular. Convolutional Neural Network (CNN) [7] has a strong ability to learn non-linear features, so it has a high accuracy in computer vision tasks such as face recognition. LFW Face dataset [8] is a well-known static face recognition dataset. Many of the face recognition methods based on deep learning achieve the highest accuracy on LFW. Hu et al. [9] proposed a new Discriminant Depth Metric Learning (DDML) method, which trains a CNN to map face to a feature space, and uses Manhattan distance measure to maximize inter-class and intra-class differences. Tran et al. [10] proposed a CNN-based three-dimensional variable face model (CNN-3DMM) method. This method inputs static face images into a trained CNN model, and obtains a three-dimensional variable face model.

Based on the analysis aforementioned, considering the shortcomings of traditional manual feature method and deep learning method. A hybrid framework for human face detection and recognition in videos is proposed in this paper. Given a frame sequence, we first use adaboost algorithm to detect
face in each frame. Then, the popular CNN model ResNet [11] is introduced to recognize the detected faces. Finally, the proposed hybrid framework is extensively validated on the YouTube Faces dataset. The experimental results show the effectiveness of the proposed framework.

2. Adaboost algorithm

The face detection algorithm based on Adaboost is proposed in 2001 by Viola Jones, which is a statistical learning algorithm [6]. The proposed algorithm makes great progress in face detection technology, and it is the mainstream face detection algorithm at present. The core idea of Adaboost algorithm is to train a few weak classifiers through an iterative training process. It is then combined into a strong classifier for detection. The whole process of Adaboost algorithm is divided into training stage and detection stage. In the training stage, a large number of positive and negative samples are trained to get a classifier. In the detection stage, the image is detected by this classifier and the face is detected. The flow chart of the Adaboost algorithm is shown in Fig. 1.

![Figure 1. The flow chart of the Adaboost algorithm](image)

Firstly, face and non-face samples are collected, and the normalized integral image of the samples is calculated. The Haar-like feature in the samples is extracted. A Haar-like feature is a separate weak classifier. The formula of weak classifier is:

\[
h(x) = \begin{cases} 
1, & p f(x) < p \theta \\
0, & \text{others}
\end{cases}
\]  

(1)

Where \(f(x)\) is the feature value of Haar-like feature; \(\theta\) indicates the threshold between positive and negative samples; \(p\) controls the direction of inequality. Comparing the eigenvalues of the detected samples with the threshold values, \(h(x) = 1\) represents the faces, and \(h(x) = 0\) represents the non-faces. After getting the weak classifier. It needs to be combined into a strong classifier. The algorithm flow of Adaboost training strong classifier is described as follows:

1) Given the sample image set \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), where \(n\) represents the number of samples, \(y_i = 0, 1\) represents the positive and negative samples respectively.

2) Initializing sample weights, \(w_{t,i} = 1/m, 1/l\), where \(m\) and \(l\) denote the number of positive and negative samples, \(m + l = n\).

3) For each iteration, \(i = 1 \ldots T\), the process is calculated as follows, where \(t\) represents the number of iterations:

   (1) Weight normalization \(w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}, i = 1, \ldots, n} \);

   (2) For each feature \(j\), training its classifier \(h_j\), calculating and selecting the minimum weighted classification error \(e_i\), as:
\[ e_i = \min e_j = \min \sum_i w_i | h_j(x_i) - y_i | \] (2)

Where the classifier \( h_i \) with the smallest weighted classification error is \( e_i \).

(3) Calculating the weight \( \alpha_t \) of weak classifier \( h_i \) by:

\[ \alpha_t = \frac{\log(1 - e_t)}{e_t} \] (3)

(4) Updating weights for each sample:

\[ w_{t+1,i} = w_{t,i} \frac{e_t}{1 - e_t} (1 - e_i) \] (4)

When the sample \( x_i \) is correctly classified, \( e_i = 0 \), vice versa \( e_i = 1 \).

(5) Finally, a strong classifier is obtained:

\[ h(x) = \begin{cases} 1, & \sum_{t=1}^{T} \alpha_i h_t(x) \geq 1 \sum_{t=1}^{T} \alpha_i \\ 0, & \text{others} \end{cases} \] (5)

Where the weight of weak classifiers is expressed as \( \alpha_t \), \( h(x) = 1 \) represents the face and \( h(x) = 0 \) represents the non-face.

3. Residual network

In view of the above situation, Deep residual networks (ResNet) introduces residual learning to solve the problem that deep network is difficult to optimize. Formally, \( H(x) \) is used to represent the optimal mapping, and the stacked non-linear layer is used to fit another mapping \( F(x) = H(x) - x \). In this case, the optimal mapping can be expressed as \( H(x) = F(x) + x \). Assuming that the residual mapping is easier to optimize than the original mapping, it is easy to push the residual to zero in extreme cases, which is much simpler than approaching another mapping. The schematic diagram of ResNet is shown in Fig. 2. \( F(x) + x \) can be represented by adding a "shortcut connection" to the feed-forward network. "Fast Connection" skips over one or more layers and performs simple identical mapping, which neither adds additional parameters nor increases computational complexity.

![Figure 2. The schematic diagram of ResNet.](image-url)
4. Experimental results and analysis

The YouTube Faces dataset contains 3425 videos collected from YouTube. According to the evaluation protocol of YouTube Faces dataset standard, 5000 video pairs were used in the experiment and divided into 10 groups, each group contains 250 positive sample pairs and 250 negative sample pairs. 100 samples were randomly selected from each video and their average similarity and processing speed were calculated. Fig. 3 shows some sample frames of YouTube Faces dataset.

![Figure 3. Some sample frames of YouTube Face dataset](image)

In order to verify the effectiveness of the feature fusion method proposed in this study, the recognition accuracy of DeepFace, Visual Geometry Group (VGG), lightened CNN and the proposed hybrid framework are compared on YouTube Faces dataset. The results show that the proposed framework in this study can improve the recognition accuracy. The high accuracy of face recognition can reach 92.50% on YouTube Faces dataset, which has good recognition effect.

| Method                                | Accuracy (%) |
|---------------------------------------|--------------|
| YouTube Faces dataset                 |              |
| DeepFace                              | 91.5         |
| Visual Geometry Group (VGG)           | 92.3         |
| lightened CNN                         | 92.0         |
| proposed hybrid framework             | 93.6         |

Table 1. Comparison of face recognition accuracy

In order to verify the effectiveness of the hybrid framework in this study for improving face recognition speed, the average processing speed is compared, and compared with VGG method and lightened CNN with higher accuracy on YouTube Faces dataset. The results are shown in Table 2, the average speed of 34 frames/s can be achieved, which fully meets the real-time processing requirements of surveillance cameras.

| Method                                | Processing speed (frame/s) |
|---------------------------------------|----------------------------|
| YouTube Faces dataset                 |                            |
| DeepFace                              | 15.3                       |
| Visual Geometry Group (VGG)           | 2.7                        |
| lightened CNN                         | 23.4                       |
| proposed hybrid framework             | 34.1                       |

Table 2. The processing speed comparison
5. Conclusion
In this paper, a video face detection recognition method based on hybrid framework is proposed. Given a frame sequence, we first use adaboost algorithm to detect face in each frame. Then, the popular CNN model ResNet is introduced to recognize the detected faces. Finally, the proposed hybrid framework is extensively validated on the YouTube Faces dataset. The framework proposed by this study has good recognition accuracy, and greatly improves the recognition speed, meets the real-time requirements in video surveillance.

References
[1] JAFRIR, ARABNIAHR. A survey of face recognition techniques [J]. Journal of Information Processing Systems, 2009, 5(2): 41-68.
[2] Kone, Jakub, Hagara M. One-Shot-Learning Gesture Recognition using HOG-HOF Features[J]. Journal of Machine Learning Research, 2017, 15(1):2513-2532.
[3] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling[C]// IEEE, International Conference on Computer Vision. IEEE, 2009:32-39.
[4] Abdelhakim A E, Farag A A. CSIFT: A SIFT Descriptor with Color Invariant Characteristics[C]// Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. IEEE, 2006:1978-1983.
[5] Lyons M, Akamatsu S, Kamachi M, et al. Coding Facial Expressions with Gabor Wavelets[C]// IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings. IEEE, 2002:200-205.
[6] Viola P, Jones M. Robust Real-time Face Detection [J]. International Journal of Computer Vision, 2004, 57(2):137-154.
[7] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection[C]// Computer Vision and Pattern Recognition. IEEE, 2015:5325-5334.
[8] HUANG G B, MATTARM, BERGT, et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments [R]. Amherst: University of Massachusetts, 2007.
[9] HU J L, LU J W, TAN Y P. Discriminative deep metric learning for face verification in the wild [C] // IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1875-1882.
[10] TRAN A T, HASSNERT, MASII, et al. Regressing robust and discriminative 3D morphable models with a very deep neural network [C] // IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1493-1502.
[11] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C] //2016 IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 770-778.