Neurocomputing 194 (2016) 10–23

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Hybrid human detection and recognition in surveillance

Qiang LIU [a,b], Wei ZHANG [a,*], Hongliang LI [c], King Ngi NGAN [b]

[a] School of Control Science and Engineering, Shandong University, China
[b] Department of Electronic Engineering, The Chinese University of Hong Kong
[c] School of Electronic Engineering, University of Electronic Science and Technology of China

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form
2 February 2016
Accepted 12 February 2016
Communicated by Shiguang Shan
Available online 19 February 2016

Keywords:
Head–Shoulder Detector
Human recognition
AdaBoost
Overlapping Local Phase Feature
Gaussian Mixture Model
Surveillance

ABSTRACT

In this paper, we present a hybrid human recognition system for surveillance. A Cascade Head–Shoulder Detector (CHSD) with human body model is proposed to find the face region in a surveillance video frame image. The CHSD is a chain of rejecters which combines the advantages of Haar-like feature and HoG feature to make the detector more efficient and effective. For human recognition, we introduce an Overlapping Local Phase Feature (OLPF) to describe the face region, which can improve the robustness to pose change and blurring. To well model the variations of faces, an Adaptive Gaussian Mixture Model (AGMM) is presented to describe the distributions of the face images. Since AGMM does not need the facial topology, the proposed method is resistant to face detection error caused by imperfect localization or misalignment. Experimental results demonstrate the effectiveness of the proposed method in public dataset as well as real surveillance video.

1. Introduction

Nowadays, surveillance cameras are deployed almost every corner and street over the world, especially in big cities, to watch and manage the activities of human being. For example, there are around 500,000 CCTV cameras in London and 4,000,000 cameras in UK [1]. It is impossible to hire enough security guard to monitor the huge number of cameras constantly, 24 h and 7 days. Generally, the camera feeds are recorded without monitoring and the videos are mainly used for a forensic or reactive response to crime or terrorism after the event happened. However, only recording surveillance video is not enough to prevent the terrorists. Intelligent detection of events and persons of interest from the camera feeds before any attack happens is urgently required for surveillance purpose.

As an intelligent surveillance system, it should be able to identify where and who is in the scene. An intelligent surveillance system mainly includes human detection and recognition. However, in practice, it is very challenging to find and recognize human when illuminations, expressions, and poses vary. Besides, surveillance videos also have low quality due to the long distance of the target from the camera, out-of-focus blur or motion blur caused by motion between the target and camera, or a combination of all factors aforementioned. Besides, camera noise and image distortion incurred by optical sensor and network transmission also affect the performance of human detection and recognition.

In the surveillance human recognition literature, most work was presented with the assumption that the face detection is given. To deal with pose variation, Gaussian mixture Models [2,3] are learned from training data to characterize human faces, head pose variations, and surrounding changes. In [4,5] use 3D model to aid face recognition to robust to facial expression and pose variations and further improvement by adding auxiliary information, such as motion and temporal information between frame images. And [6] uses “Frontalization” face to do face recognition and gender estimation. Ma et al. [7] improved the accuracy of pose estimation by investigating the symmetry property of the face image. To deal with the illumination variations, Thermal Infrared Sensor (TIRS) [8] was used to measure energy radiations from the object, which is less sensitive to illumination changes. However, thermal images have low resolution and are unable to provide rich information of the facial features. To account for blurring problem, Hennings-Yeomans et al. [9] first performed restoration to obtain images with better quality [10], and then fed them into a recognition system. Rather treating restoration and recognition separately, Zhang et al. [11] proposed a joint blind restoration and recognition model based on sparse representation to deal with frontal and well-aligned faces. Grgic et al. [12] also provided a surveillance face database collected in uncontrolled indoor environment using five types surveillance cameras of various qualities and applied principal component analysis (PCA) for face recognition. In [13,14], each face was described in terms of multi-region modelled by probabilistic distributions, such as GMMs, followed by a normalized

* Corresponding author.
E-mail address: davidzhangsdu@gmail.com (W. ZHANG).

http://dx.doi.org/10.1016/j.neucom.2016.02.011
0925-2312/

distance calculated between two faces, which can be efficient to deal with faces with illumination and misalignment. However, face recognition is still an open problem in surveillance, although techniques [15–17] used in face recognition literature [18–20,74] perform well with the cooperative subjects in controlled applications. Also, current face detectors are unable to find the face well in the low-quality surveillance video.

In this paper, we present a hybrid human recognition system by integrating face detection and recognition together as shown in Fig. 1. For face detection, we propose to find the Head–Shoulder (HS) region first by the Cascade Head and Shoulder Detector (CHSD), and then employ the trained body model to get the face region for recognition. In face recognition, to represent face region discriminatively, we propose an Overlapping Local Phase Feature (OLPF) which is robust to image blur and pose variation without adversely affecting discrimination performance. To model faces robustly, a Fixed Adaptive Gaussian Mixture Model (FGMM) is developed to describe the distribution of the face data, but FGMM may be degraded because of different subjects needing different numbers of Gaussians to model the variations of faces. Therefore, an Adaptive Gaussian Mixture Model (AGMM) is proposed to optimally build the model for each subject. Without face topology, the proposed AGMM is insensitive to the initial face detection without alignment. Combining AGMM and OLPF, our method can handle faces with multiple uncontrolled issues in surveillance, such as misalignment, pose variation, illumination changing, and blurring. The proposed detection and recognition scheme can be extended to other objects of interest with similar properties such as cars and animals.

The organization of the paper is arranged as follows: In Section 2, we give the structure of CHSD and the details of how to train each filter in the CHSD. The proposed face recognition algorithm are discussed in Section 2.1. Extensive experiments are given in Section 2.2 to demonstrate the robustness of our method. Conclusions are summarized in Section 2.2.1.

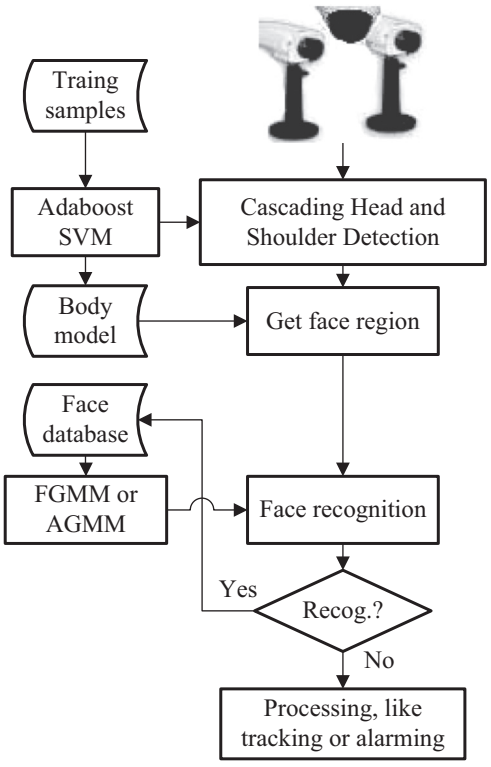

Fig. 1. Diagram of proposed system.

2. Cascade Head and Shoulder Detection

As aforementioned, in general surveillance condition, people and the target scene cannot be strictly controlled. The face to be recognized may not appear as assumed in [11,19], such as the frontal face with proper lighting. So the captured faces may differ substantially in pose, illumination and expression. Some examples are given in Fig. 2 from an indoor surveillance application to show the variations of pose and illumination in the face region. For these cases, traditional face detector [17,21] may not work well to locate the face region effectively and correctly. To overcome these problems in unconstrained conditions, we propose to detect HS region first, and then use the human body model to obtain the face region.

The proposed method is inspired by [22,23] with the use of a dense grid of Histograms of Oriented Gradients (HoG) and linear Support Vector Machine (SVM) to detect human. However, we found that those detectors are not enabled to allow fast rejection in the early stages. It works slowly and can only process 320×240 images at 10 frame per second (fps) in a sparse scanning manner. In this paper, we intend to speed it up to real-time without quality loss by cascading new classifiers.

The idea of CHSD is to use a cascade of rejecters to filter out a large number of non-HS samples while preserving almost 100% of HS regions. Thus the number of candidates can be reduced significantly before more complex classifiers are called upon to achieve low false positive rates. As shown in Fig. 3(a), CHSD includes three parts: initial feature rejecter, Haar-like rejecter, and HoG classifier.

2.1. Initial feature rejecter

In this rejecter, one of the features is the regional variances which can be obtained by limited computations[1] from two integral images, i.e., integral image and integral image of the squared image. Those integral images will also be used to perform illumination normalization in the preprocessing step and feature calculation in the Haar-like rejecter, so no additional computation is required in this rejecter. Assuming that σ_k denotes the variance of the kth region, our training process is described in Algorithm 1.

The other feature of the first rejecter is the difference between two blocks no matter whether they are adjacent or not. The training method in Algorithm 2 is similar to that in Algorithm 1 with a few minor modifications from steps (a)–(c).

Algorithm 1. Training for rejecter using variance features.

1. Input training data $(\langle x_1, y_1\rangle, \ldots, \langle x_n, n\rangle)$ where $y_i \in \{0, 1\}$ for non-HS and HS regions,respectively.
2. Initialize rejecter label $l_i = 0$, for $y_i = 0$;
3. For $t = 1, \ldots, T$:
   a. Find the minimal and maximal values of σ_k for each region k from the training samples, which are denoted by ${\sigma_k}^{\min}$ and ${\sigma_k}^{\max}$, respectively.
   b. Compute the rejection number r_k for non-HS training samples, with a parity p adjusting the in-equality direction: $r_k^p = \sum_{y_i = 0, l_i = 0} sign\,|\,p\sigma_{i,k} > p\sigma_k^p\,|$, $p = -1$ for ${\sigma_k}^{\min}$ and $p = 1$ for ${\sigma_k}^{\max}$
   c. Choose the region with the highest rejection number
   d. Set label $l_i = 1$ for all rejected sample $\{i\}$.
4. Output the combined classifiers.

[1] Any two-rectangle feature can be computed in six array references, any three-rectangle feature in eight, and any four-rectangle feature in just nine.