MULTI-VIEW FACE DETECTION OVER MULTI-CAMERA SURVEILLANCE SYSTEMS

Ching-chun Huang (黃敬群), Jay Chou (周節), and Sheng-Jyh Wang (王聖智)

ABSTRACT

In this paper, we propose a multi-view face detection system that locates head positions and indicates the direction of each face in 3-D space over a multi-camera surveillance system. To locate 3-D head positions, conventional methods relied on face detection in 2-D images and projected the face regions back to 3-D space for correspondence. However, the inevitable false face detection and rejection usually degrades the system performance. Instead, our system searches for the heads and face directions over the 3-D space using a sliding cube. Each searched 3-D cube is projected onto the 2-D camera views to determine the existence and direction of human faces. Moreover, a pre-process to estimate the locations of candidate targets is illustrated to speed-up the searching process over the 3-D space. In summary, our proposed method can efficiently fuse multi-camera information and suppress the ambiguity caused by detection errors. Our evaluation shows that the proposed approach can efficiently indicate the head position and face direction on real video sequences even under serious occlusion.

Keywords: Face detection; Multi-view face detection; Multi-camera surveillance system; Image information fusion;

1. INTRODUCTION

Up to now, a lot of algorithms have already been proposed to solve the face detecting problem. The most popular approach is the training-based approach which collects lots of face data to construct a database for training. With the face database, a suitable classifier is learned to detect faces with high detection rate and low false alarm rate. For example, Viola and Jones [1] proposed the Adaboosting detection algorithm which is fast, robust and reliable to detecting frontal faces in 2-D images. Nowadays, several algorithms with similar structures have been proposed to improve the accuracy of detection based on Adaboosting detection algorithm. However, there still exist many difficulties in face detection, one of which is the detection of non-frontal faces. For non-frontal face detection, there appear view-dependent deformation and variation. Hence, these frontal face classifiers usually cannot be directly applied to non-frontal face detection.

In many applications, such as visual surveillance system, human faces in the captured images may not be upright and frontal. In these cases, the detection of faces becomes much more complicated. These non-frontal faces usually contain less information and present more diversity. This fact makes non-frontal detection a lot more sensitive to noise, background, illumination, and facial model.

A few methods for multi-view face detection have been proposed in recent years. They could be roughly divided as single-camera systems and multi-camera systems. For single-camera systems, Huang et al.’s [2] method provided an important reference. In their system, they proposed a method to construct a rotation invariant multi-view face detector. Their method was composed of a Width-First-Search tree detector structure, a Vector Boosting algorithm for learning strong classifiers, a domain-partition-based learning method, sparse features in granular space, and a heuristic search for sparse feature selection. Their system can detect multi-view faces with low computational complexity and high detection accuracy. However, the detection task may fail in some cases, such as low-resolution faces, inter-object occlusions, and incomplete human faces in images. It is also difficult for the method to detect the back side of human heads. Apparently, non-frontal face detection based on a single view of observation would be very difficult. The use of multiple cameras may somewhat relieve the difficulties in non-frontal face detection.

Multi-camera systems could provide us more information about the scenarios we concern. For a multi-camera surveillance system, more than one camera is installed within a surveillance zone. These cameras are installed at different locations, capture more information of the targets, and help users to monitor the targets in a more precise and efficient way. In [3], Huang and Wang proposed an efficient way to fuse 2-D foreground detection result from multi-camera views. In their system, they adopted a probabilistic method to label multiple targets based on a Markov network. In [4], Zhang et al. presented a system that integrates temporal
and spatial information to build a multi-camera multi-view face detection system inside a room. By integrating temporal and spatial information with the dynamic programming approach, they aimed to detect the face of the lecturer for a lecture scenario inside an appropriately equipped smart room. In their approach, the multi-view face detector was implemented based on the FloatBoost approach [5]. However, those methods usually detect the face regions on 2-D images and then project the regions back to 3-D space for locating the 3-D positions. The inevitable false face detection and rejection may degrade the system accuracy even under a multi-camera surveillance environment.

In our system, we aim to establish multi-view face detection for an intelligent multi-camera surveillance system. Here, we plan to accomplish a system that is capable of detecting all targets’ faces within the surveillance zone and is able to indicate the direction of each face inside a surveillance zone. Unlike most frameworks doing detection in 2-D images, our goal is to do this job in 3-D space since it would help us to well use the 3-D geometry knowledge such as the size of human face, the rough height of a human head above the ground plane, and etc. In detail, our system searches for the targets over the 3-D space using a sliding cube. Each searched 3-D cube is projected onto the 2-D camera views to determine the existence and direction of human faces. With the 3-D geometry prior, we could detect faces on 2-D images without trying different scales of patch sizes if comparing with many previous methods. Moreover, our approach can efficiently combine 2-D information from different camera views and suppress the ambiguity caused by 2-D detection errors. By fusing information form multi-camera views, we can infer the location of faces and their directions.

The rest of this paper is organized as follows. In section 2, we present the main idea of the proposed framework. In section 3, we explain how we estimate the locations of candidate targets on the 3-D ground plane. In section 4, we detail our Multi-view face detection framework for locating the head positions and extracting the face directions. Experimental results and discussions are presented in section 5. Last, Section 6 concludes this paper.

2. OVERVIEW

2.1. System Overview

In this paper, the whole system operates on an environment where a surveillance zone is monitored by multiple cameras. The main goal of our face detection framework is to locate human heads and detect the face directions. Our system includes two steps – (1) 3-D position estimation and (2) multi-view face detection framework - as shown in Fig 1.

For the first step, we detect the locations of candidate targets. Here, we fused multi-view foreground detection results in order to identify the positions of candidate targets on the 3-D ground plane. The goal of this step is to filter out most impossible positions in the 3-D space and to speed up the searching process in the second step. For the second step, we aim to locate the optimal head position and determine the face direction of each target in the 3-D space. Here, we search the face with a 3-D cube within the possible subspaces, which is determined by the first step. Next, the 3-D cube at a possible face location is projected onto the 2-D camera views to get projected image regions. These image regions are verified by using eight pre-trained face detectors, which correspond to eight face views, in order to measure the probabilities for different face views. The measured probabilities from multiple images are finally fused in a systematic manner. Based on the fused probabilities, the finding of head positions and face directions are finally formulated as an optimization problem and solved in a unified way. In the following sections, we will explain the details of each functional block in our system flow.

Fig. 1: Flow chart of the proposed multi-view face detection framework.

3. 3-D POSITION ESTIMATION

3.1 Background Subtraction on a Single Camera

To identify the location of candidate targets on the 3-D ground plane, we need to detect the foreground regions which is accomplished by taking the difference between the current image and the reference background in a pixel-wise manner. Here, we model the reference background based on the Gaussian mixture model (GMM) approach [6]. An example of the foreground detection result is shown in Fig. 2. Note that the foreground regions are neither perfectly silhouetted nor well connected due to the influence of noise, variation of illuminations, and shadows.

Fig. 2: (a) The original image. (b) The result after background subtraction.
In (1), \( X \) represents a location \((x_1, x_2)\) on the ground plane. \( N \) is the number of static cameras in the multi-camera system. \( F_i \) denotes the foreground image of the \( i \)th camera view. Assume the size of camera views is \( M_i \times N_i \). The point \((m, n)\), which is in the range of \( 0 \leq m \leq (M_i - 1) \) and \( 0 \leq n \leq (N_i - 1) \), denotes the coordinates of a pixel on the foreground image. Then \( F_i \) is defined as

\[
F_i(m, n) = \begin{cases} 
1 & (m, n) \in \text{foreground regions} \\
0 & (m, n) \notin \text{foreground regions} 
\end{cases}
\] (2)

Moreover, given the location \( X \), we assume the foreground images are conditionally independent of each other. Also, we assume the prior \( p(X) \) is uniform distributed that indicates the equal possibility of finding a moving person at \( X \). Therefore, (1) can be rewritten as

\[
p(X) p(F_1, \ldots, F_N | X) = p(X) \prod_{i=1}^{N} p(F_i | X). \] (3)

On the other hand, to formulate \( p(F_i | X) \), we approximate a moving target at the ground position \( X \) as a rectangular pillar. The height \( H \) and radius \( R \) of the pillar are modeled as independent Gaussian random variables, with their Gaussian priors \( p(H) \) and \( p(R) \) being pre-trained via training samples. Based on the pre-calibrated projection matrix of the \( i \)th camera and a sample pair \((H,R)\), we project the pillar onto the \( i \)th camera view to get the projected image \( M_i \). Here we define \( M_i \) on the \( i \)th camera view as

\[
M_i(m, n \mid H, R, X) = \begin{cases} 
1 & (m, n) \in \text{projected regions} \\
0 & (m, n) \notin \text{projected regions} 
\end{cases}
\] (4)

The expectation of the overlapped region of \( M_i \) and \( F_i \) with perspective normalization offers a reasonable estimate about \( p(F_i | X) \). That is, we define \( p(F_i | X) \) as

\[
p(F_i | X) = \int \int \Omega_i(H, R, X) p(H) p(R) dH dR,
\] (5)

where the normalized overlap correlation, \( \Omega_i \), is defined as

\[
\Omega_i(H, R, X) = \frac{\int \int F_i(m, n) M_i(m, n \mid H, R, X) dmdn}{\int \int M_i(m, n \mid H, R, X) dmdn}
\] (6)

Based on (3) and (5), the TDP distribution could be calculated. An example of TDP distribution is shown in Fig. 4. Here, we may find the TDP is composed of many clusters; each cluster indicates a candidate target on the ground. Therefore, the candidate targets can be identified by some clustering algorithms, such as Mean-Shift clustering. After clustering, we can extract the number of candidate targets \( N_T \) inside the current surveillance zone and estimate the ground location \( X_i \) for the \( i \)th target by finding its corresponding cluster centers. Please refer to [3] for more details.

**4. MULTI-VIEW FACE DETECTION FRAMEWORK**

![Fig. 3: (a) The illustration of the model-driven approach for information fusion. (b) A pillar at a true location generates larger overlapped regions. (c) A pillar at a wrong location generates smaller overlapped regions.](image)

![Fig. 4: (a) Input images. (b) The TDP of four moving targets in the surveillance zone.](image)
Fig. 14: (a) Detection result at a correct 3-D position. (b) Detection result at an incorrect 3-D position. (c) The blue line corresponds to the likelihood values of eight hypotheses of face orientations at the correct position, and the green line corresponds to the likelihood values of eight hypotheses at the incorrect position.

We also show some detection results in Fig. 15. To clearly present our outputs, we use bounding boxes with different colors to indicate different targets. We also mark the detected face direction onto the bird-eye view of the surveillance zone. In this example, there are two persons in the scene. As shown in the figure, our system can detect faces and identify the face directions even if some serious occlusion occurs or someone is out of image view. In Fig. 15(a), there is an occlusion case in the top-left image and there is a missing person in the lower-right image. For this example, our system can still find the approximate locations of the faces and the face directions, as shown in Fig. 15(b). Another experimental result is illustrated in Fig. 15(c-d).

6. CONCLUSION

In this paper, we present a multi-view face detection system over a multi-camera surveillance system. Through this system, we can detect all target faces in the given images and identify the direction of each face in the 3-D space. Unlike existing approaches whose performance are usually degraded by inter-object occlusion, the proposed system does not directly detect targets over the 2-D image domain nor project the 2-D detection results back to the 3-D space for correspondence. In our system, we search for the targets over small cubes in the 3-D space. Each searched 3-D cube is projected onto the 2-D camera views to determine the existence and the direction of a human face. With this approach, we can efficiently combine 2-D information from different camera views to make a more reliable and robust inference even under the inevitable false face detection and rejection of our 2-D classifiers.

REFERENCES