J. Vis. Commun. Image R. xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Visible-light and near-infrared face recognition at a distance ☆

Chun-Ting Huang [a,*], Zhengning Wang [b], C.-C. Jay Kuo [a]

[a] Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA
[b] Department of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 4 February 2016
Revised 9 August 2016
Accepted 24 September 2016
Available online xxxx

Keywords:
Face recognition
Cross-distance matching
Cross-environment matching
Two-stage filtering
Image restoration
Locally Linear Embedding (LLE)

ABSTRACT

A method to solve the problem of face recognition at a distance (FRAD) under the visible-light (VIS) and the near-infrared (NIR) spectra is presented in this work. For images taken under visible light at day time, we perform the coarse-scale alignment/enhancement to eliminate a set of unlikely candidates at the first stage. Then, the fine-scale alignment/enhancement steps are conducted to refine the candidate list furthermore iteratively at the second stage. To address the additional challenge associated with NIR images captured at night time, we incorporate a restoration mechanism that reconstructs low-quality patches through a locally linear embedding (LLE) process with a local constraint. It is shown by experimental results that our FRAD solution outperforms state-of-the-art methods on both VIS and NIR images.

1. Introduction

With a rapid growth of security demand, authorities in many countries adopt video surveillance systems to monitor possible threats for public security and enforce traffic management. For instance, London has installed about half-million cameras to observe public spaces, and major cities in the United States such as New York and Los Angeles are extending their surveillance networks largely. As the result, the fast increasing volume of large video data has become harder to manage. Besides, incidents like Boston Marathon bombing require heavy manual labor to examine low-quality face images within the surveillance camera's footage. The method to improve efficiency and accuracy of automatic face recognition in the context of video surveillance imposes a major challenge in research community.

Even though there have been progressive developments in automatic face recognition, most state-of-the-art methods focus on faces with variant poses yet at a close distance with sufficient quality. The captured face by a surveillance system can be of low resolution and poor quality, which is degraded by an uncontrolled outdoor environment such as long distance at day or even night time. This leaves a noticeable gap between the clear image within datasets and the real footage obtained by surveillance systems. For example, a great majority of face recognition researchers study faces with an arbitrary pose such as those in the Labeled Faces in the Wild (LFW) dataset [1]. In contrast, little research has been conducted to solve the problem of face recognition at a distance (FRAD). Usually, long distance outdoor face images and short distance indoor face images for matching are called *probe* and *gallery* sets, respectively, in this context.

Since most surveillance footages taken at a long distance can be easily distorted by illumination and polarization, the recognition performance is severely affected. Normalization is a critical step in handling distortions caused by the long distance. It can be categorized into two types: geometrical normalization and photometrical normalization. They are called face alignment and face enhancement, respectively, in this work.

Under normal lighting circumstances, face alignment includes finding an initial rough face shape reference, and approaching the ground truth via iterative optimization. Facial landmark localization is a well known iterative technique in finding the coordinates of essential components. Cascaded regression is a regression-based method proposed for landmark localization. It was introduced by Dollar et al. [2] for pose estimation in image sequences. Later, it was applied to face alignment. Cao et al. [3] proposed a regression method with two-stage training, where cascaded regression was extended to the context of an affine transform. Xiong and De la Torre [4] applied cascaded regression with the SIFT feature and examined the derived solution from a gradient descent view. It is called the supervised descent method (SDM). Yan et al. [5] adopted a similar framework with the "learn-to-rank" and the "learn-to-combine" modules placed before and after

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: chuntinh@usc.edu (C.-T. Huang).

http://dx.doi.org/10.1016/j.jvcir.2016.09.012
1047-3203/

the main alignment module, respectively. Automatic face alignment techniques and systems have been extensively tested. For instance, Wagner et al. [6] used the spare representation in their alignment algorithm for the Multi-PIE dataset. Geng and Jiang [7] developed an automatic alignment system based on both holistic and local features and conducted experiments on the AR, GT and ORL datasets. Deng et al. [8] proposed a transform-invariant PCA (TIPCA) method to achieve automatic alignment, and tested its performance on the FERET dataset. Recently, deep learning [9] and 2D/3D alignment [10,11] offer competitive performance in this field.

Most automatic alignment algorithms mentioned above target at unconstrained faces such as different face poses, expressions or even occlusion. Faces in a long distance outdoor environment are distorted due to blurring, illumination and various weather conditions. Ban et al. [12] proposed a face alignment method to address distortions caused by blurring and low resolution. However, since their dataset was built in a short-distance (1–3 m) indoor environment, their alignment problem is different from that occurs in the long-distance outdoor environment.

As compared with geometrical face normalization, photometric face normalization has received much less attention. A system that incorporates wavelet decomposition, deblurring, denoising and linear stretching was proposed in [13] to recover quality loss due to the long distance. The reported recognition performance ranges from 50% to 70% due to low image resolution and quality. Another face enhancement technique rooted in retinex theory was proposed by Land and McCann [14]. It examined relative lightness (instead of absolute lightness) in a local region to mimic the human visual experience. Furthermore, Land [15] presented a lightness computation method. By following their work, Jobson et al. proposed a single-scale retinex (SSR) method in [16] and extended it to a multi-scale retinex (MSR) method in [17]. Rahman et al. [18] proposed MSR with color restoration (MSRCR) to handle illumination variation.

An ideal surveillance system should operate around the clock, including both day and night time. Currently, cameras equipped with flash lights are used for night time to offer acceptable performance. However, they are not appropriate for long distance or convert surveillance. As a result, we need to consider other options for night time face recognition. Methods like near infrared (NIR), shortwave infrared (SWIR), and thermal infrared have been studied in the literatures. NIR has become popular in recent years for several reasons [19]. First, NIR is not visible to human eyes, and it is desired to capture face expressions without interrupting subjects in acquisition. Second, the environmental factor has less impact to NIR when compared with others. Third, the NIR illuminator can penetrate glasses easily, which provides additional information if the test subject wears glasses. Generally speaking, NIR offers a good choice for night time long distance face recognition.

The FRAD problem under VIS and NIR spectra is a challenging topic. A two-stage alignment/enhancement filtering (TAEF) system was proposed for VIS FRAD problem in [20]. It contains alignment and enhancement in the coarse-scale stage and facial matching with a refined candidate pool in the fine-scale stage. In this work, we address the FRAD problem under both VIS and NIR spectra as an extended version of [20]. The material on NIR FRAD problem in this work is completely new. To bridge the gap between VIS and NIR, we propose a restoration system that significantly improves the quality of enhanced NIR face images as compared with prior art in [21]. The restoration system is developed using the Locally Linear Embedding (LLE) method [22] that reconstructs image patches learned from two manifolds. It preserves image's local characteristics by constraining the reconstruction process within a certain region so that the recovered information will be independent of other regions.

The rest of this paper is organized as follows. Cross-distance, cross-environment and cross-spectral datasets are reviewed in Section 2. We present a system to address the cross-distance and cross-environment facial matching problem under the visible light in Section 3. Then, we examine the facial matching problem with the NIR images as the input, and propose a restoration mechanism to relate NIR and VIS images in Section 4. The performance of the proposed solution is evaluated by extensive experimental results including state-of-the-art deep learning approach in Section 5. Error cases are analyzed in Section 6 for additional insights into the proposed solution. Finally, concluding remarks and future works are given in Section 7.

2. FRAD datasets

There are few publicly available datasets collected for the FRAD problem. The study began with only the cross-distance scenario (VIS-to-VIS) with the UTK-LRHM dataset [13] in 2008. It contains 55 subjects with distances ranging from 10 to 16 m in an indoor environment and 48 subjects with distances from 50 to 300 m in an outdoor environment. Another FRAD dataset was built by Rara et al. [23] for sparse-stereo reconstruction in 2009. It has 30 subjects with three distances (namely, 3, 15, and 33 m). Tome et al. [24] evaluated the effect of distance degradation for several matching methods based on the "Face still dataset" of the NIST multiple evaluation grand challenge (MBGC) [25].

Researchers have started to consider both the cross-distance and the cross-spectral challenges for FRAD datasets since 2011. The near-infrared face recognition at a distance (NFRAD) dataset was built by Maeng et al. [26]. It has 50 subjects with both VIS and NIR photos in a controlled environment, where the captured distances are 1 m and 60 m. This dataset provides some pose change (the frontal view, the slight left and the right face view angles). However, since the NIR illuminator generates a halo-like light pattern around the subject at the 60 m distance, its usage becomes very limited. The second dataset, collected by Bourlai et al. [27], has 103 subjects captured at 30, 60, 90 and 120 m in the NIR outdoor environment as the probe set, and 5 feet in the VIS indoor controlled conditions as the gallery set, respectively. This dataset does not cover outdoor VIS images. Finally, the LDHF dataset [28] provides 100 subjects with three standoff distances: 60, 100, 150 m VIS and NIR outdoor probe images, and 1 m VIS and NIR gallery images with better quality comparing with other two datasets. The LDHF dataset is the only one that is available in the public domain with both cross-distance and cross-spectral characteristics.

Four exemplary LDHF visible-light images are shown in Fig. 1, which are taken at 100 and 150 m, VIS and NIR, respectively. They have the same image resolution (i.e., 5184×3456 pixels) but different face sizes (i.e., 220×220 pixels for the 100-m image and 120×120 pixels for the 150-m image on average). The illumination in the LDHF dataset can be roughly categorized into three types: normal, foggy, and back-lighted. The two images in the first row of Fig. 1 demonstrate how image quality can be affected by the foggy and the back-lighted environments. Apparently, both face alignment and enhancement techniques are needed before facial matching.

Experimental results have been conducted and reported for the LDHF dataset recently. Maeng et al. [26] applied Gaussian smoothing and histogram equalization as the preprocessing step and extracted the dense scale invariant feature transform (DenseSIFT) [29] from 32×32 overlapping patches. Then, each patch was divided into 4×4 grids, where an 8-bin gradient orientation histogram was calculated to form a 128-D feature vector. The matching distance between feature vectors of two VIS images

Please cite this article in press as: C.-T. Huang et al., Visible-light and near-infrared face recognition at a distance, J. Vis. Commun. (2016), http://dx.doi.org/10.1016/j.jvcir.2016.09.012