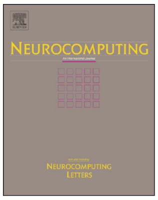
Neurocomputing 207 (2016) 48–59

Contents lists available at ScienceDirect

Neurocomputing

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



The THU multi-view face database for videoconferences and baseline evaluations[☆,☆☆]

Xiaoming Tao[a], Linhao Dong[b,*], Yang Li[a], Jianhua Lu[a,b]

[a] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[b] *School of Aerospace Engineering, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Article history:
Received 25 September 2015
Received in revised form
9 March 2016
Accepted 28 March 2016
Communicated by Xu Zhao
Available online 10 May 2016

Keywords:
Face video database
Videoconferences
Gaze-correction
Multi-view
Evaluation protocol

ABSTRACT

In this paper, we present a face video database and its acquisition. This database contains 31,500 video clips of 100 individual figures from 20 countries. The primary purpose of building this database is to serve as a standardized test video sequences for any research related to videoconferences, such as gaze-correction and model based face reconstruction, etc. To be specific, each of the subjects was filmed by 9 groups of synchronized webcams under 7 illumination conditions, and was requested to complete a series of designated actions. Thus, face variations including lip shape, occlusion, illumination, pose, and expression are well presented in each video clip. Compared to the existing databases, the proposed THU face database provides multi-view video sequences with strict temporal synchronization, which enables evaluations on current and future possible gaze-correction methods for conversational video communications. Besides, we discuss the evaluation protocol based on our database for gaze-correction, where three well-known methods were tested. Experiment results show that, under this evaluation protocol, comparisons of performance can be obtained numerically in terms of peak signal-to-noise ratio (PSNR), demonstrating the strengths and weaknesses of these methods under different circumstances.

1. Introduction

1.1. Background and motivation

Recently, video communication over Internet has attracted much attention from the public. Its unique features make it irreplaceable in many scenarios, such as teleconference [1], telemedicine [2], video calling, etc. Also, its related design of multicast protocol, rate allocation, and household application have been well investigated [3–5], providing accessible solutions to home-based video communications. However, when it comes to user experiences, the current video communication applications often fall short of satisfaction. For instance, one of the most glaring incongruities often found in personal video calling is that natural eye contact cannot be maintained since it is impossible to stare at the camera and the conversation window simultaneously.

To address this problem, several schemes [6–9] were proposed to correct gaze direction for users at both ends. They are generally categorized into *image-based* [6,7] and *model-based* [8–10] schemes. For image-based schemes, calibrated stereo cameras are usually applied to capture images from two separate views. Utilizing binocular vision processing, a virtual central view is synthesized; therefore the gaze direction is corrected accordingly. Differing from image-based schemes, model-based ones typically require a depth map from either a stereo camera set [8] or a depth camera [9] to build a 3D face model. In [10], authors applied a generic 3D head mesh model to fit the features extracted from individual face images. By using morphing techniques, the face model can be rotated in the correct direction, whereas texture-mapping is performed to transfer the texture patches in the original image onto the direction-corrected face model. These schemes, however, were only able to be evaluated qualitatively and subjectively, each in isolation, by observing the output synthetic sequences. There lacked a unified criteria with which the schemes can be compared quantitatively. Therefore, a specialized database is desirable in order to evaluate the performance of these gaze-correction methods.

[☆] This work was supported by the National Basic Research Project of China (973) (2013CB329006), National Natural Science Foundation of China (NSFC, 61101071, 61471220, 61021001), and Tsinghua University Initiative Scientific Research Program (20131089365).

[☆☆] Part of this work was published in proceedings of IEEE ICIP'15, Québec City, Canada.

[*] Corresponding author.

E-mail addresses: taoxm@tsinghua.edu.cn (X. Tao), linhaodong@tsinghua.edu.cn (L. Dong), liy-11@mails.tsinghua.edu.cn (Y. Li), lhh-dee@tsinghua.edu.cn (J. Lu).

http://dx.doi.org/10.1016/j.neucom.2016.03.055
0925-2312/© 2016 Elsevier B.V. All rights reserved.

1.2. Face database and its design

In the field of automatic face analysis and perception, significant progress has been made in several sub-domains such as facial expression analysis, face tracking, 3D face modeling, etc. [28–30]. From a recognition viewpoint, these schemes usually involve learning-based solutions that require training datasets; their performances are often directly related to the size and the quality of the available training sets. Therefore, diversity and comprehensiveness are important factors for a database. Since mid-1990s, several institutions have begun building face databases in various forms, including images [11–21,23–27], videos [11,15,16,22,24] and audio-visual sequences [11,15,22].

Among these databases, early works including M2VTSDB [11], Yale Face Database [12], and AR Face Database [13] pioneered diverse methodologies in design. In M2VTSDB (Release 1.00), synchronized video and speech data as well as image sequences were recorded for the research of identification strategies. A variety of facial variations such as hairstyles, accessories, etc. were presented. Also, 3D information can be obtained from the head movements in video sequences. As an extended version of M2VTSDB, XM2VTSDB [15] collected data by using digital devices with a larger sample space for more reliable and robust training. However, the illumination conditions in both of these databases are constant and ideal, therefore the impact of illumination on face identification cannot be effectively evaluated. In the Yale Face Database and the AR Face Database, various controlled illuminations were introduced, such that the impact of illumination conditions on the same face can be well studied. Both the Yale Face Database and the AR Face Database, nevertheless, present only frontal face images, precluding the ability to test the success rate of face recognition algorithms on profile faces.

It has been reported that the variations between the images of the same face caused by illumination and viewing direction are likely larger than that caused by change in face identity [31]. Hence, face images in databases like the Yale Face Database B [19], the PIE Database [20], Multi-PIE [25], and the CAS-PEAL Face Database [23] were taken under several designated poses and controlled illuminations. In the Yale Face Database B, a geodesic dome consisting of 64 strobes was constructed to offer 45 illumination conditions for training purpose; meanwhile, images of the same face were taken in 9 viewing directions to provide a possibility of building 3D models. However, the facial expression was not considered as one of the variations. In both PIE and Multi-PIE Databases, pose, illumination and expression were variables for each face; the geometric and color calibration information for each camera was recorded as well. Specifically, more deliberate facial expressions were contained in Multi-PIE. With similar acquisition methods to PIE, the CAS-PEAL Face Database provides a large sample space of Chinese people as a supplementary for a specific ethnicity. Considering the challenges of face recognition in surveillance scenario, COX Face Database [24] offers a large video-based database of 1000 subjects acquired by 3 camcorders under uncontrolled light conditions with low spatial resolution to simulate the practical surveillance environment. It can also serve as a diverse training set with different views of face in front of complex backgrounds. Table 1 lists some popular face databases with their respective features.

While the aforementioned databases have been successfully applied in several fields, they were not constructed specifically for videoconference or conversational video communication scenarios; specifically, none offers synchronized videos with multiple views. Aiming at providing a video database for videoconferences to supplement the currently available corpus, we have collected 31,500 raw videos from 100 volunteers, the total size of which is around 5 TB. Our database is specially designed for conversational videos; thereby, the scenes in the video clips simulate real-life conversation scenarios. To be specific, each subject was requested to complete a series of designated actions that may appear during one video communication, including speaking, head movements, drinking with a cup, reading, and performing one facial expression; through these actions variations of lip shape, occlusion, pose, and expression were exhibited. Moreover, each subject was filmed under 7 different controlled illumination conditions. Multiple cameras were arranged in groups for video acquisition. For each clip, 5 out of 13 available cameras were used for simultaneous recording. Such simultaneous multi-view recording enables us to propose an evaluation protocol for gaze-correction methods, with which the three typical methods including two image-based methods [6,7], and one model-based method [8] were evaluated as examples to give baseline performances. The evaluation results show quantitative comparisons among these methods different illuminations and action scenarios.

The remainder of this paper is organized as follows. The hardware setup is described in Section 2, where the arrayed camera system and illumination arrangements are explained. Post-processing, including geometric and color calibration for the cameras are described in Section 3. In Section 4, the design and structure of this database is introduced with examples. The proposed evaluation protocol and the baseline experiment results are presented in Section 5. Finally, we draw the conclusion in Section 7.

2. Hardware setup

This section describes the hardware setup for video acquisition in our studio. In Section 2.1, the layout and function of the webcam array is introduced. In Section 2.2, the detailed arrangements of various illumination conditions are described.

2.1. Multi-view camera array

To record multi-view videos for one face, 13 Logitech® c310 HD webcams are used, each of which is able to capture 1280 × 720 HD image sequences. A rectangular camera rack carrying all the cameras was custom designed to mimic the dimension of a regular 16:9 19-in monitor. As shown in Fig. 1, the 4 cameras labeled as U, R, D, and L are called *acquisition cameras*, since one or more of them can be physically presented in a conversational video communication system, and actually used by gaze-correction methods; the cameras labeled as C1 to C9 (sequenced in the zig-zag order) are called *reference cameras*, since they do not actually appear on screen in a real system, and are used to capture the target view which the gaze-correction methods must attempt to synthesize. They are placed across the interior of the rack to signify the fact that the actual messaging window may not necessarily appear at the center of the monitor.

During each round of the video collection, only 1 out of 9 reference cameras is active, along with cameras U, R, D, and L always recording. In other words, 5 cameras are required to record videos simultaneously. Since one computer can support at most 3 webcams to record videos of 1280 × 720 @ 25 fps at the same time due to bandwidth limit of the USB port, these 5 cameras are controlled by two inter-connected computers (3+2, separately) in strict temporal synchronization. In case of unexpected difference of frame rate happening among cameras, two timestamps are made in the form of .yml file after each recording, where the absolute moment of each frame is written. By using such timestamps, frames from 5 cameras can be trimmed accordingly. Timestamps are stored along with the video clips. While recording, the subject is asked to perform the same sequence of actions under each illumination condition and for each reference camera,