TUNIR: A Multi-Modal Database for Person Authentication under Near Infrared Illumination

SHUYAN ZHAO, RALPH KRICKE, ROLF-RAINER GRIGAT
Hamburg University of Technology
Vision Systems, E-2
Harburger Schloßstr. 20, 21079 Hamburg
GERMANY

Abstract: Audio- and video-based person authentication has gained significant attention. Common databases are crucial for comparative evaluation of algorithms. Almost all of the publicly available databases contain images/videos captured in visible illumination. In this paper we present an acquisition system and procedure for collection of a multi-modal database under near infrared (NIR) illumination. Audio and video data of 74 subjects have been recorded under simulated realistic conditions. Multi-view images with different facial expressions have been captured under strictly controlled conditions. Ground truths, i. e. the manually marked eye positions and lip corners, are also provided. The database is publicly available for research purposes on request from Hamburg University of Technology (`http://www.ti1.tu-harburg.de/TUNIR`).

Key–Words: Infrared, Person Authentication, Multi-Modal Database, Face Recognition, Lipreading

1 Introduction

Biometric person authentication has received ever-increasing attention in research and industry due to high demand on security recently. Authentication performance is generally improved by combining multiple modalities, for example faces and speech. Empirical evaluation of algorithms needs to be carried out on standardized databases. Availability of public databases is significant for advancement in this field. As far as we know, there are four multi-modal databases including faces and voice publicly available: the M2VTS database [2], the extended M2VTS database [6], DAVID-BT database [5] and the recent Banca database [3]. Face images/videos of these databases were all captured using visible imagery. Numerous evaluations prove that illumination is one of the challenges of face recognition. One way to address the illumination problem is to use imagery other than that in visible band, which is nearly invariant to changes in ambient illumination. For example, Equinox infrared face database [1] contains images captured using a long-wave infrared in the spectral range of $8 - 12$ μm. (For some subjects additional short-wave infrared ($0.9 - 1.7\mu$m) and middle-wave infrared ($3 - 5\mu$m) data are available.) However, thermal infrared is not desirable compared with NIR, because of the higher cost of thermal sensors and poorer quality of the thermal images.

This paper puts forward a multi-modal (audio and video) database under active NIR illumination. We call it TUNIR database. In Section 2, we introduce the hardware configuration. In Section 3 acquisition and content of the data are described. Section 4 presents ground truth labeling. We conclude this paper in Section 5.

2 Hardware Configuration

The infrared (IR) band covers a wavelength range of electromagnetic spectrum from 780 nm to 1 mm. Usually the wavelength range of $780 - 3000$ nm is called near infrared (NIR). What this work is interested in is NIR from 780 nm to 1100 nm. The benefits of the use of NIR are:

- It enables face detection and recognition to work under variable visual illumination condition.
- Common cameras with silicon sensors, sensitive to wavelengths less than 1100 nm, can be used. Compared with thermal sensors, silicon sensors are inexpensive and capture images of high quality.
- The *bright pupil effect*, generated under the NIR illumination condition, can be employed to assist eye detection [4, 10].

We use the Philips PCVC840K webcam with a SONY ICX098BQ CCD sensor[7]. The diagonal

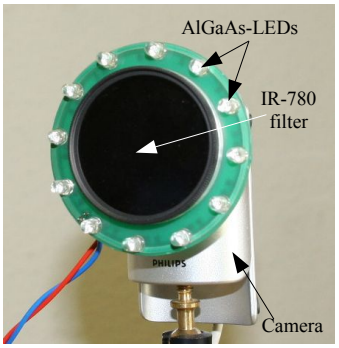


Figure 1: Camera and NIR illumination

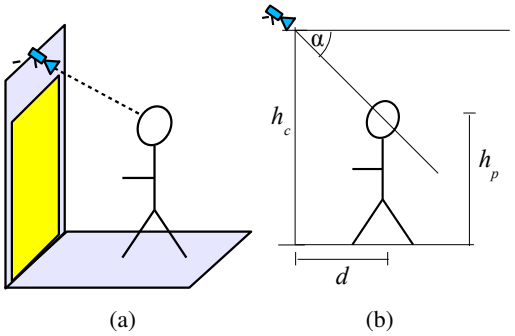


Figure 2: Camera setup

length of the sensor is 4.5 mm. The lens of the camera has a focal length of 6 mm, its aperture has a diameter of 3 mm. The videos are recorded at a frame rate of 30 fps in progressive mode. The resolution is set to 320×240. Gain and white balance are determined by the camera automatically. The audio data are captured using the built-in microphone of the webcam.

Figure 1 illustrates the camera and illumination system. We use 12 AlGaAs-LEDs (875 nm), distributed on a circle around the camera axis, to illuminate the scene. Note that we have removed the built-in infrared blocking filter of the camera. To eliminate variations in ambient illumination, we have added a daylight cut-off filter IR-780.

The camera is mounted horizontally centered above the entrance door, at a height of $h_c = 1.95$ m with a tilt angle of $\alpha = 30°$ (see Figure 2).

3 Data Collection

Audio and video data were collected using the above configuration in two phases. In total 74 people from Europe, Asia, and Africa participated in the recording. The participants stood at a distance $0.39\text{ m} \leq d \leq 0.78$ m away from the camera. Their eyes were located at a height of $1.56\text{ m} \leq h_p \leq 1.86$ m. They were asked to look at the camera, but their head positions were not controlled. Thus head movements occurred during recording.

Figure 3: Snapshots of video sequences of Phase I

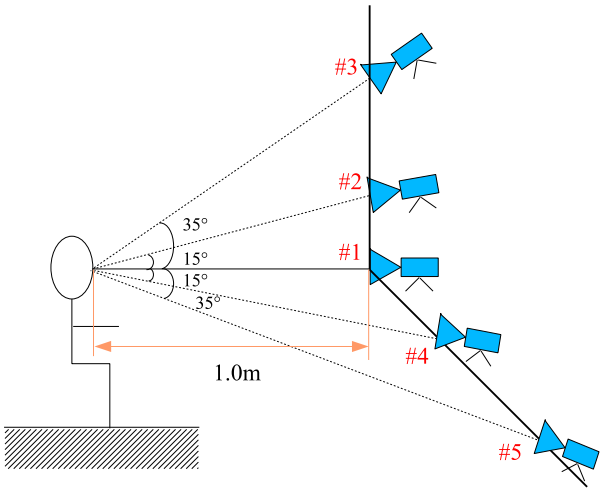


Figure 4: Schematic illustration of setup for acquisition of multiview images

The speech content is digits, i. e. "zero, one, two, three, ..., nine". Audio data are stored in PCM format, while videos are stored in H.263.

Data Collection: Phase I 10 subjects were involved and two sequences of each of them were recorded. Figure 3 illustrates snapshots of some video sequences. Since facial expression analysis and pose detection are important for face recognition, still images with various facial expressions and head poses were also recorded. Figure 4 shows the setup used in this session. The view angles were strictly controlled. Camera 1 acquires frontal face images, Camera 2 captures top view at 15°, Camera 3 top view 35°, while Camera 4 and 5 capture side views at 15° and 35° respectively. The variation of facial expressions were simulated by the mimic generated when people speak vowels a, e, i, o, and u in German. Two images were captured for each vowel. The variation of facial expressions and poses can be seen in Figure 5.

During this recording session, the appearance variation, such as eyeglasses and beard were also considered, since it has a significant effect on face recog-

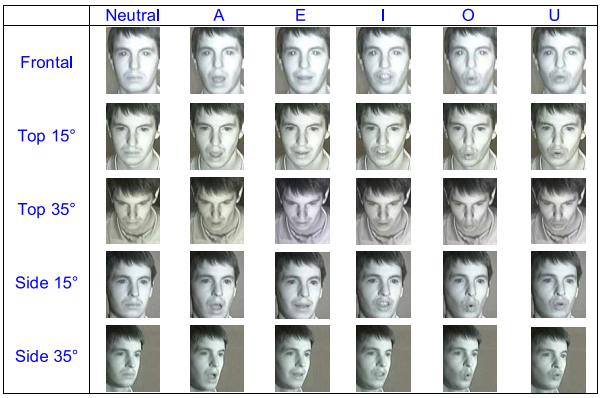


Figure 5: Variation of facial expressions and poses

Figure 6: Variation of appearance: glasses

nition. Figure 6 and Figure 7 show some examples.

Data Collection: Phase II 64 subjects different from those in phase I participated. To simulate realistic applications, video sequences captured the whole procedure: the subject is approaching, standing in front of the camera and speaking, then leaving (see Figure 8). Four sequences of each person were acquired, two sequences without glasses, and two with glasses. Each sequence contains about $200 - 300$ frames.

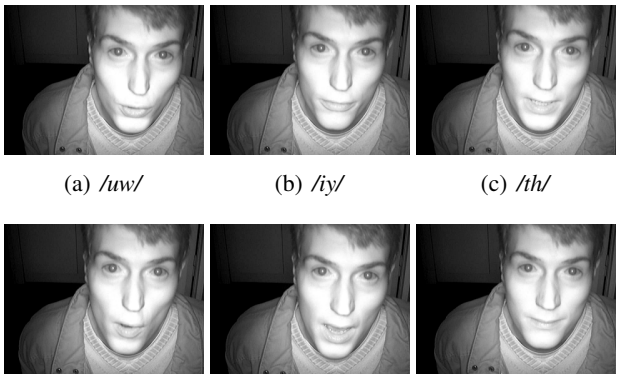
4 Labelling of the Data

Our proposed method for a person authentication system [9] relies on accurate localization of the eyes in

Figure 7: Variation of appearance: beard

Figure 8: Sample frames of one video sequence: the person approaching, speaking, and leaving.

Figure 9: Snapshots of video sequences of Phase II



(a) */uw/* (b) */iy/* (c) */th/*

(d) */ao/* (e) */ay/* (f) */sil/*

Figure 10: Images of one subject showing different viseme classes

the image. To be able to quantitatively measure the performance of the eye localization method, we have marked the eye positions in selected frames manually.

For each sequence of Phase I, 50 consecutive frames have been selected and labeled. For each of these images one text file including the coordinates of the eyes and lip corners has been saved.

Data from Phase II are used to evaluate features for visual speech reading (lip reading) operating on NIR. Neti et al. [8] suggested a mapping of 44 phonemes to 13 distinguishable mouth shapes (visemes). We have selected 6 visemes from this list, namely */uw/* of the word "two", */th/* as well as */iy/* of "three", */ao/* of "four", */ay/* of "nine", and additionally a closed mouth */sil/* corresponding to silence. Each of these visemes are visible in 2-5 consecutive frames. These frames have been selected for all 4 recording sessions of 16 subjects. This builds our viseme database of 1113 images. For each of these images, the coordinates of the left and right eyes as well as the left and right lip corners have been manually marked, which again are stored in a text file. Images showing these visemes can be found in Figure 10.

5 Conclusion

The TUNIR face database provides the research community with the possibility to test their algorithms for audio- and video-based person authentication under near infrared illumination. So far audio and video data from 74 subjects have been recorded. Since under active near infrared illumination the reflection of eyeglasses is an issue for face detection and recognition, this problem is specially taken into account. Among the four sessions of each individual, two are with glasses and two without glasses. Although the database has been acquired in the laboratory environment, the recording condition is uncontrolled in order to simulate a "real world " scenario. Ground truths, namely positions of eyes and lip corners, as well as six distinct viseme classes have been manually collected. To assist research of pose and expression invariant face recognition, still images also have been captured. We hope that our database offers a platform for research institutes to compare their algorithms for multimodal person authentication under near infrared illumination.

Acknowledgements: This work is part of the project *KATO – Innovative Cabin Technologies* funded by the German Federal Ministry of Economics and Technology.

References:

[1] Equinox infrared face database. `http://www.equinoxsensors.com/products/HID.html`.

[2] The M2VTS database. `http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html`.

[3] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The banca database and evaluation protocol. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[4] A. Haro, M. Flickner, and I. Essa. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2000.

[5] J. S. D. Mason, F. Deravi, C. C. Chibelushi, and S. Gandon. Digital audio visual integrated database: Final report. Technical report, Department of Electrical and Electronic Engineering, University of Wales Swansea, 1996.

[6] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *International Conference on Audio- and Video-based Biometric Person Authentication*, 1999.

[7] David Molyneaux. Web cameras for astronomy imaging, July 2005. `http://homepage.ntlworld.com/molyned/web-cameras.htm`.

[8] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Technical Report WS00AVSR, Johns Hopkins University, CLSP, 2000.

[9] Faisal Shafait, Ralph Kricke, Islam Shdaifat, and Rolf-Rainer Grigat. Real time lip motion analysis for a person authentication system using near infrared illumination. In *2006 IEEE International Conference on Image Processing*, pages 1957–1960, Oct 2006.

[10] S. Zhao and R. Grigat. Robust eye detection under active infrared illumination. In *International Conference on Pattern Recognition*, Hong Kong, 2006.