# The CMU Face In Action (FIA) Database

Rodney Goh[1], Lihao Liu[1], Xiaoming Liu[2], and Tsuhan Chen[1]

[1] Electrical and Computer Engineering Department,
Carnegie Mellon University, Pittsburgh 15213, USA
{rhgoh, llihao, tsuhan}@andrew.cmu.edu
[2] GE Global Research, one research circle, New York 12308, USA
liux@research.ge.com

**Abstract.** Our team has collected a face video database named the CMU Face In Action (FIA) database. This database consists of 20-second videos of face data from 180 participants mimicking a passport checking scenario. The data is captured by six synchronized cameras from three different angles, with an 8-mm and 4-mm focal-length for each of these angles. We performed the collection in both a controlled, indoor environment and an open, outdoor environment for each participant. Our data collection was taken in three sessions over a period of ten months. We aimed for a three month separation between sessions for each participant. We expect the database to be useful for analysis and modeling of faces and gestures.

## 1 Introduction

There are many existing databases containing face images under controlled conditions, FERET [1], CMU PIE [2], ORL [3], Yale Database [4], UMIST [5]. However, as more and more researchers begin working on video-based face recognition as opposed to traditional image-based face recognition, there is a greater demand for a database of human faces in video sequences. With such a database, the benefits of video-based face recognition can be explored.

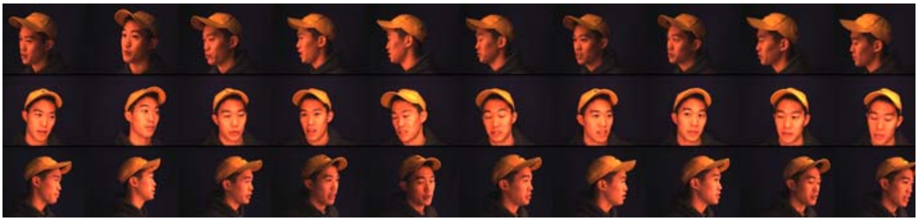
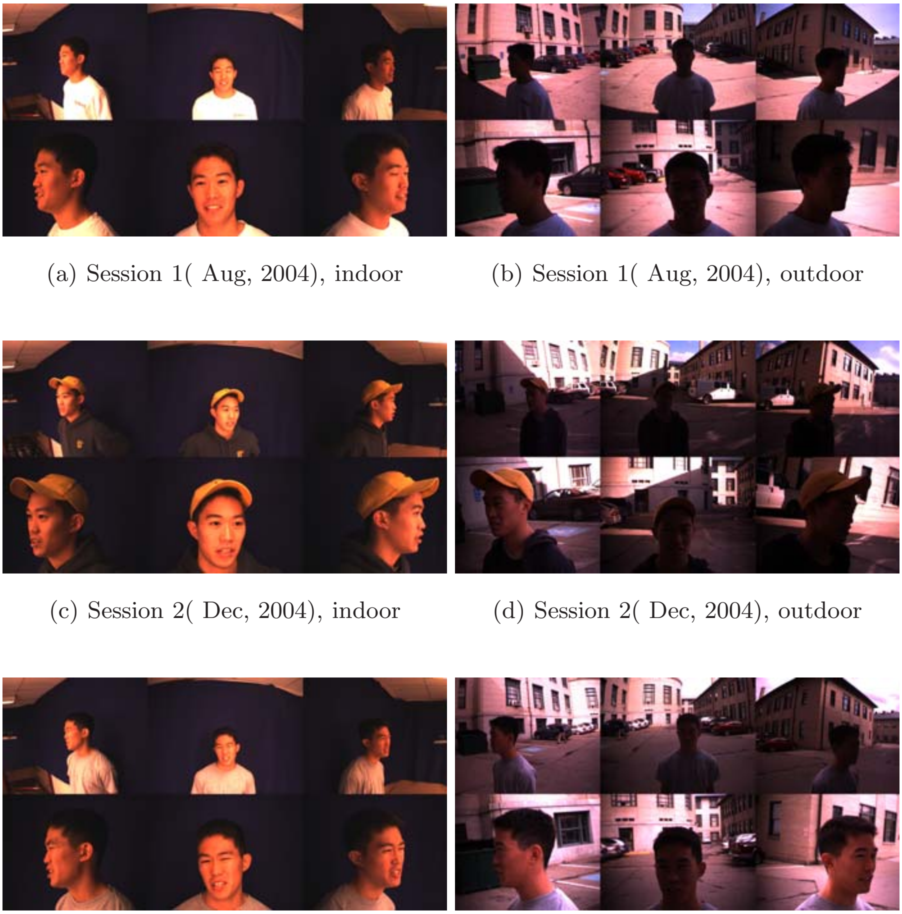


**Fig. 1.** Sample images of the Face In Action database taken from the three 8-mm focal-length cameras. Here we show 10 intermittent frames among the 600 JPEG images to give a clear idea of our face video data.

W. Zhao, S. Gong, and X. Tang (Eds.): AMFG 2005, LNCS 3723, pp. 255–263, 2005.

We have collected such a face video database, calling it CMU Face In Action (FIA) database (Figure 1). The collection was performed for 180 participants in both indoor and outdoor environments, three times per participant, over a ten month period. We captured our videos from three different angles, with two different focal-lengths for each angle. For each of these 20-second videos, participants were asked to mimic a passport checking scenario, providing a large range of gestures, facial expressions, and motions.



(a) Session 1( Aug, 2004), indoor

(b) Session 1( Aug, 2004), outdoor

(c) Session 2( Dec, 2004), indoor

(d) Session 2( Dec, 2004), outdoor

(e) Session 3( Apr, 2005), indoor

(f) Session 3( Apr, 2005), outdoor

**Fig. 2.** Frame number 300 of 600 for the indoor (left) and outdoor (right) scenarios from all three sessions. User-dependent gestures and expression variations are expected.

## 2 Database Variation

In face database collection, one samples the face in multiple dimensions, such as pose, illumination, expression, aging, etc. In our CMU FIA capturing system, we sampled in the following dimensions: motion, pose, image resolution, illumination and variations over time.

Motion and pose were left participant-dependent. In order to vary image resolution, we utilized two focal lengths for each angle from which we captured. To sample variations over time, we conducted our data collection in three sessions, aiming for three months separation between sessions. Illumination was varied by our two different environments: a controlled, indoor environment and an open, outdoor environment. The indoor environment was fixed with a blue background and fluorescent lighting. The outdoor environment utilized natural lighting, that could be affected by season and climate as seen as Figure 2. The variables that remained constant between these environments were the data rate and quality of the videos, camera angles, procedure, and the face of each participant. The poses for each sequence will have varied, as they are dependent to the participant. The sequences from the outdoor scenario can be used to study how well a video-based face recognition system performs in a natural setting.

### 2.1 Equipment

For our video collection, we utilized six OEM-style IEEE-1394 board level Dragonfly cameras (Figure 3) from Point Grey Research Inc. [6], along with their corresponding synchronizing mechanisms. The cameras utilize the Sony ICX424 sensor, which has a maximum resolution of 640x480 pixels and 24-bit true color images (RGB), and a maximum frame rate of 30 Hz. The cameras were put on adjustable arms of two identical aluminum carts, one used in our indoor scenario and the other for our outdoor scenario. One IEEE-1394 bus can process a data stream from a maximum of three cameras, based on the data rate of 640x480 pixels at 30 frames per second. Therefore, we utilized two separate buses, each responsible for synchronizing three cameras. Since we were aiming to synchronize

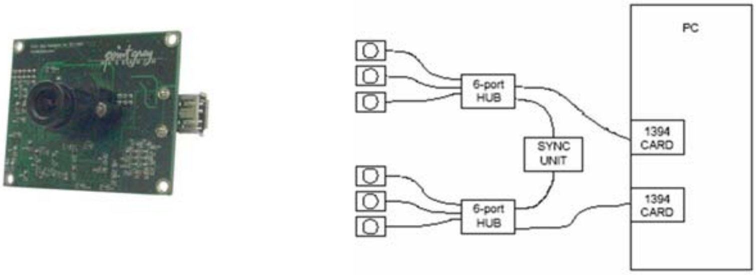


**Fig. 3.** The Dragonfly camera from Point Grey Research Inc. (left). The Synchronization Unit (right) plays the role of synchronizing two IEEE-1394 buses.

all six cameras together for our database, and only three cameras may be synchronized on one bus, we attached the two buses, each processing three cameras, into a separate camera synchronization unit, which then serves to synchronize the two groups together. Based on our experiences, the speed of the hard drive, rather than the CPU speed, is the bottleneck of the capturing system. We used more memory as cache to compensate for the latency of the hard drive. Even so, we occasionally encountered “out of sync” frames. For the purpose of our database, out of a total 600 frames, we made sure to keep this number of “out of sync” frames to under 60 in the indoor scenario, and under 150 in the outdoor scenario. In most cases, we had no “out of sync” frames.

### 2.2 Positioning

As shown in Figure 4, we have built a cart for mounting the capturing system. There are six cameras on the C-shape arm whose height can be adjusted manually from 1.5 meters to 1.7 meters according to the different height of each participant. All of the cameras point to the same center spot from a distance of 0.83 meters. We placed a red cross mark on the floor at this center spot as a reference point for our participants. Since the C-shape arm can be adjusted vertically by the linear bearing according to the height of the subject, the face is essentially captured by three pairs of cameras with the same vertical angle, but three different horizontal angles (-72.6 , 0 , 72.6 ) respectively. The six cameras were arranged into three pairs. For each pair of cameras, one camera was set to an 8-mm focal-length, which results in a face area of around 300 x 300 pixels, and the other to a 4-mm focal-length, which results in a face area of around 100 x 100 pixels. The video sequence with larger face area can be used for applications demanding high-resolution face images, such as 3D reconstruction, while the sequence with smaller face area presents face data closer to the size used in video surveillance applications and gesture analysis. Two carts were used for

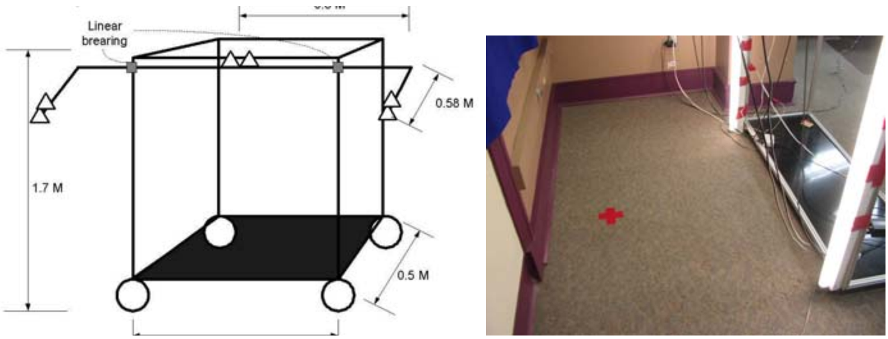


**Fig. 4.** A simple diagram of our cart to show exact dimensions and the location of our cart

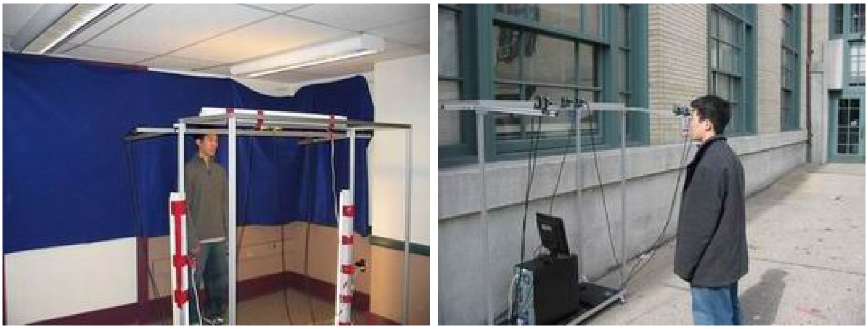
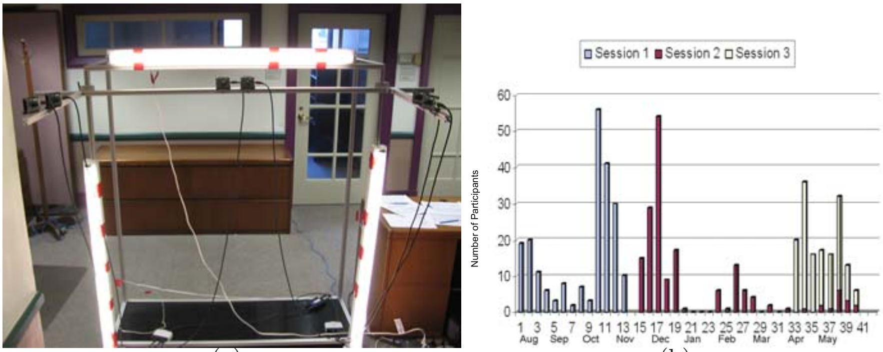


Fig. 5. Snapshots of data collection in progress for the indoor (left) and outdoor (right) scenarios

capturing our data, one in a controlled indoor environment, and the other in an open outdoor environment, as seen in Figure 5. The indoor cart was always kept in the same location, whereas the outdoor cart, for concern of security and weather damage, was moved for each day that data was collected. Therefore, the location of the outdoor cart may have varied by a few feet day to day, on the occasion that the view was blocked by parked cars in the area. The red cross reference and cart were adjusted accordingly.

## 2.3 Illumination

In the indoor scenario, we had a controlled environment with a blue background made from a felt material, and fluorescent bulb lighting. In addition to standard overhead, office lighting, a 40 inch, 40 watt Phillips "Soft White" fluorescent bulb was affixed on either side of the cart, as well as to the top of the cart,



(a) (b)

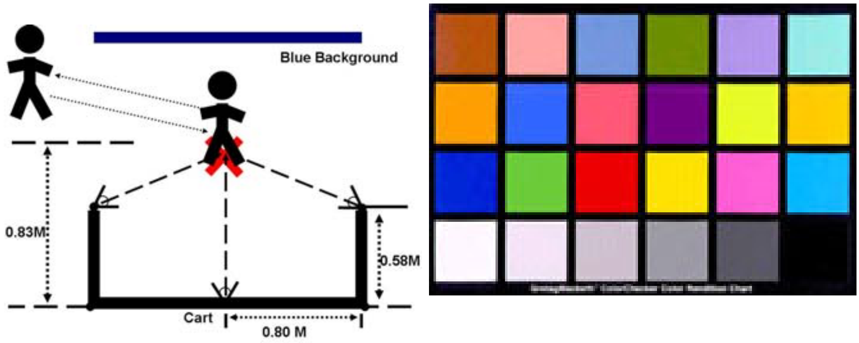
Fig. 6. (a) The positioning of the fluorescent lighting for the indoor scenario and (b) Our collection schedule represented in a bar graph, separated into sessions by color pointed directly forward from the cart (Figure 6 (a)), in order to provide sufficient lighting for our video. In the outdoor scenario, we utilized natural daylight lighting, that was dependent upon the variant weather and season. The data collection did not take place during heavy precipitation, in order to save our equipment from water damage.

### 2.4 Time Variance

We captured data in three sessions, each session occurring during a different season: Late summer/fall, winter, and spring. Our goal was to have a three month separation between sessions for each participant. The weekly schedule for all three sessions is shown in Figure 6 (b). We began collecting data for session one in August, 2004. Session two began in the beginning of November, 2004. Session three began in the middle of March, 2005. The total collection spanned 10 months, or 40 weeks. The overlap between sessions two and three are attributed to a large amount of precipitation and cold weather during the months of session two, delaying our data collection for that session. The specific dates of each participants' collections has been documented.

## 3 Collection Procedure and Calibration

To simulate real-world face motion and gestures, we asked each of the participants to mimic a passport checkpoint at an airport, as shown in Figure 7 (a). Each participant was asked to begin from the side of the cart, walk in and stand on the red cross mark in front of the the cart, and simulate the gestures and conversation typical of a passport check. After about 20 seconds, the participant was asked to exit back off to the side. Each session consisted of two similar runs, one taking place with the indoor scenario and one in the outdoor. There was no audio recorded.



(a) (b)

**Fig. 7.** (a) A diagram illustrating participants' procedure and (b) ColorChecker Color Rendition Chart

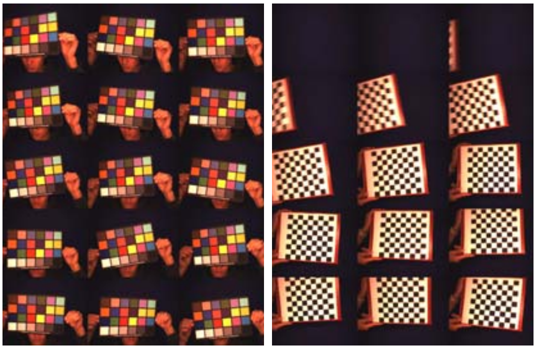


**Fig. 8.** Sample images of color (left) and camera (right) calibration data. Here we show 15 intermittent frames for the front camera.

In order to ensure color consistency among all the cameras, daily color calibration data was taken. We waved a GretagMacbeth$^{TM}$ [7] ColorChecker Color Rendition Chart (Figure 7 (b)) in the area where a participant's face would be. This chart measures 8.5 x 11 inches and contains an array of 24 scientifically designed colored squares. Six cameras captured the continuously moving chart simultaneously. The color calibration data was taken at 20 frames per second for 5 seconds. Figure 8 shows sample images of our color calibration data. This color calibration data was taken for each day that we collected face data.

We also captured camera calibration data for each subject, using a 9x9 black and white checkerboard. Similarly to the color chart, the checkerboard was waved at the location that was clearly visible from all cameras. We captured the calibration data at 7 frames per second for 6.5 seconds. Also, Figure 8 shows sample images of our camera calibration data. The color calibration and camera calibration data were only collected in the indoor scenario.

## 4 Database Organization

### 4.1 Data

We used the maximum resolution and data rate available to us through the Dragonfly cameras, which is 640x480 pixels at 30 frames per second. Each video sequence is 20 seconds long. Our video sequences, which are taken in raw .AVI format, are converted into raw .PGM images for every frame captured. Giving consideration to the large amount of disk space required, we converted the frames to JPEG format with 90% quality. Each image consumes about 100 kb of hard disk space. Thus the total required storage space for our database is ≈ 380GB for only the .JPEG files. We must also put into consideration the space required for the daily color calibration data and the camera calibration data for each subject. Daily color calibration takes up ≈ 5 GB and camera calibration takes up ≈ 15 GB. Thus, the total storage for the participants' images, daily color calibration, and participants' camera calibration is 400 GB. The raw .AVI files require a space of 1.02 GB per participant, leading to an additional required space of 180 participants, 1100 GB. Therefore, the total required disk space for our complete database is 1100GB+400GB = 1500 GB.

### 4.2 Demographics

We captured the first session of CMU FIA for 214 participants, 180 of whom returned for the second session, of whom 153 returned for a final, third session. Of these participants for the first session, 38.7% are female and 61.3% are male. The youngest participant is 18 years old. The oldest participant is 57 years old. The mean age of the participants is 25.4 years old. In addition, we recorded whether or not each participant wears glasses and/or has facial hair. This data, along with gender and age, is also documented.

## 5 Possible Usage

The CMU FIA database, with different kinds of variations such as pose, illumination, expression, aging, and etc. is beneficial to the task of recognizing human faces. The CMU FIA database is especially helpful to pose and face gesture variation related research, which is the most difficult to model [8]. Given the variety of variation we were sampling in the CMU FIA database, we suggest that CMU FIA can be used in the following studies:

- Video-based face recognition.
- Pose invariant face recognition [9].
- Three-dimensional face reconstruction from multiple views or from a video sequence.
- Face recognition with respect to image resolution.
- Outdoor illumination invariant face recognition.
- Face recognition over periods of time.
- Face and facial gesture modeling and analysis.

## 6 Summary and Availability

We have collected a face video database named CMU FIA database. By mimicking the passport checking scenario, six synchronized cameras capture human faces simultaneously from three different poses. We have performed the collection in both indoor and outdoor environments three times, in order to capture the face variance over time. Our data is open to all those interested in using it for research purposes. Those interested may contact Tsuhan Chen (tsuhan@cmu.edu) and send us one or several hard drives. The disk space for only the JPEG images is about 400 GB. The raw .AVI files require an additional 1100 GB.

## References

1. P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss, "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results". *Army Research Lab technical report 995*, October 1996.
2. Terence Sim, Simon Baker, and Maan Bsat, "The CMU Pose, Illumination, and Expression Database". *IEEE Transactions on Pattern Analysis and Machine Intelligence , Vol. 25, No. 12 pages 1615 - 1618*, December 2003.
3. ORL Face Database. http://www.uk.research.att.com/facedatabase.html
4. Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman , "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose". *IEEE Transaction Pattern Analysis Machine Intelligence, Vol. 23, No 6, pages 643 - 660*, 2001.
5. Daniel B. Graham, and Nigel M. Allinson, "Characterizing Virtual Eigensignatures for General Purpose Face Recognition". *recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman- Soulie, and T. S. Huang, editors, Face Recognition: From Theory to Applications, Vol. 163 of NATO ASI Series F, Computer and Systems Sciences, pages 446 - 456*, 1998.
6. Point Grey Research Inc. http://www.ptgrey.com/
7. GretagMacbeth AG. http://www.gretagmacbeth.com/
8. Xiaoming Liu, and Tsuhan Chen, "Online Modeling and Tracking of Pose-Varying Faces in Video". *Video Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition* , June 2005.
9. Xiaoming Liu, and Tsuhan Chen, "Video-based Face Recognition Using Adaptive Hidden Markov Models". *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition* , Madison, Wisconsin, June 2003.