

# **The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces**

**Terence Sim, Simon Baker, and Maan Bsat**

**CMU-RI-TR-01-02**

## **Abstract**

Between October and December 2000 we collected a database of 41,368 images of 68 people. By extending the CMU 3D Room we were able to image each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions. We call this database the CMU Pose, Illumination, and Expression (PIE) database. We hope to increase the number of subjects over the next few months and to image as many people as possible for a second time. In this technical report we describe the extensions we made to the CMU 3D Room, the collection procedure, and the organization of the database.

# 1 Introduction

People look very differently depending on a number of factors. Perhaps the three most significant factors are: (1) the pose; i.e. the angle at which you look at them, (2) the illumination conditions at the time, and (3) their facial expression; i.e. whether or not they are smiling, etc. Although several other face databases exist with a large number of subjects [Philips *et al.*, 1997], and with significant pose and illumination variation [Georghiades *et al.*, 2000], we felt that there was still a need for a database consisting of a fairly large number of subjects, each imaged a large number of times, from several different poses, under significant illumination variation, and with a variety of facial expressions.

Between October 2000 and December 2000 we collected such a database consisting of 41,368 images of 68 subjects. We call this database the CMU Pose, Illumination, and Expression (PIE) database. To obtain a wide variation across pose, we used 13 cameras in the CMU 3D Room [Kanade *et al.*, 1998]. To obtain significant illumination variation we augmented the 3D Room with a “flash system” similar to the one constructed by Athinodoros Georghiades, Peter Belhumeur, and David Kriegman at Yale University [Georghiades *et al.*, 2000]. We built a similar system with 21 flashes. Since we captured images both with the background lighting on and with it off, we obtained 43 different illumination conditions. Finally, we asked the subjects to pose with several different “expressions.” In particular, we asked them to give a neutral expression, to smile, to blink (i.e. shut their eyes), and to talk. (These are probably the four most common “expressions” in normal life.)

We did not capture images of every person under every possible combination of pose, illumination, and expression because of the huge amount of storage space required. There are also technical difficulties in simultaneously capturing illumination variation and expression

variation if the expression changes over time, as for example during talking. The database therefore consists of two major partitions, the first with pose and expression variation only, the second with pose and illumination variation; i.e. there is no simultaneous variation in illumination and expression. In the remainder of this report, we describe the capture setup in the 3D Room, the capture procedure, and the organization of the database.

## 2 Capture Apparatus and Capture Procedure

Obtaining images of a person from multiple poses requires either multiple cameras capturing images simultaneously or multiple “shots” taken consecutively (or a combination of the two.) There are a number of advantages of using multiple cameras: (1) the process takes less time, (2) if the cameras are fixed in space, the (relative) pose is the same for every subject and there is less difficulty in positioning the subject to obtain a particular pose, (3) if the images are taken simultaneously we know that the imaging conditions (i.e. incoming illumination, etc) are the same. This final point can be useful for detailed geometric and photometric modeling of objects. On the other hand, the disadvantages of using multiple cameras are: (1) we actually need to possess multiple cameras, digitizers, and computers to capture the data, and perhaps more importantly, (2) the cameras need to be synchronized. The shutters must all open at the same time and we must know the correspondence between the frames in the multiple video streams output by the cameras and captured in the computers.

Setting up a synchronized multi-camera imaging system is therefore quite an engineering feat. Fortunately such a system had already been built at CMU, namely the CMU 3D Room [Kanade *et al.*, 1998]. So, we simply reconfigured the 3D Room and used it to capture multiple images of each person simultaneous from multiple (fixed) poses.

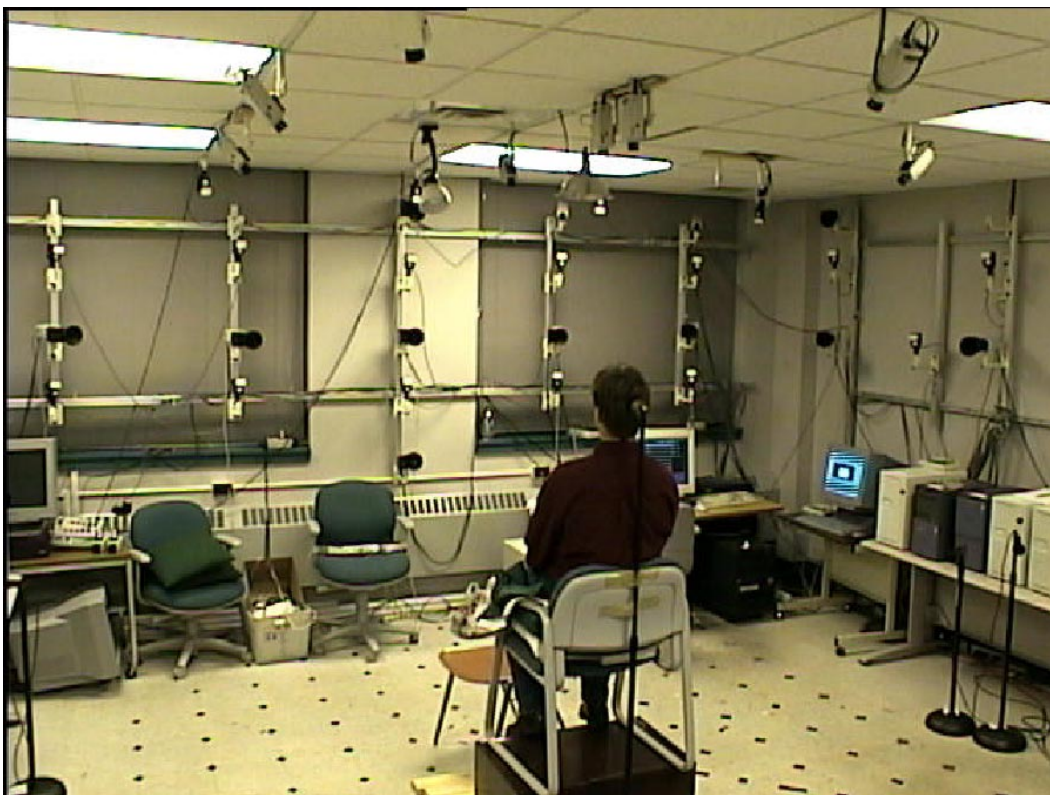


Figure 1: The capture setup in the 3D Room [Kanade *et al.*, 1998]. The subject sits in a chair with their head in a fixed position. We used 13 Sony DXC 9000 (3 CCD, progressive scan) cameras with all gain and gamma correction turned off. We augmented the 3D Room with 21 Minolta 220X flashes controlled by an Advantech PCL-734 digital output board, approximately duplicating the Yale “flash system” used to capture the database described in [Georghiades *et al.*, 2000].

## 2.1 Setup of the Cameras in the 3D Room: Pose

In Figure 1 we include an image of the capture setup in the 3D Room. In the current version of the 3D Room there are 49 cameras, 14 very high quality (3 CCD, progressive scan) Sony DXC 9000’s, and 35 lower quality (single CCD, interlaced) JVC TK-C1380U’s. We decided to use only the Sony cameras so that the image quality is approximately the same across the entire database. Due to other constraints we were only able to use 13 of the 14 Sony cameras, however, this still allowed us to capture 13 poses of each person simultaneously.



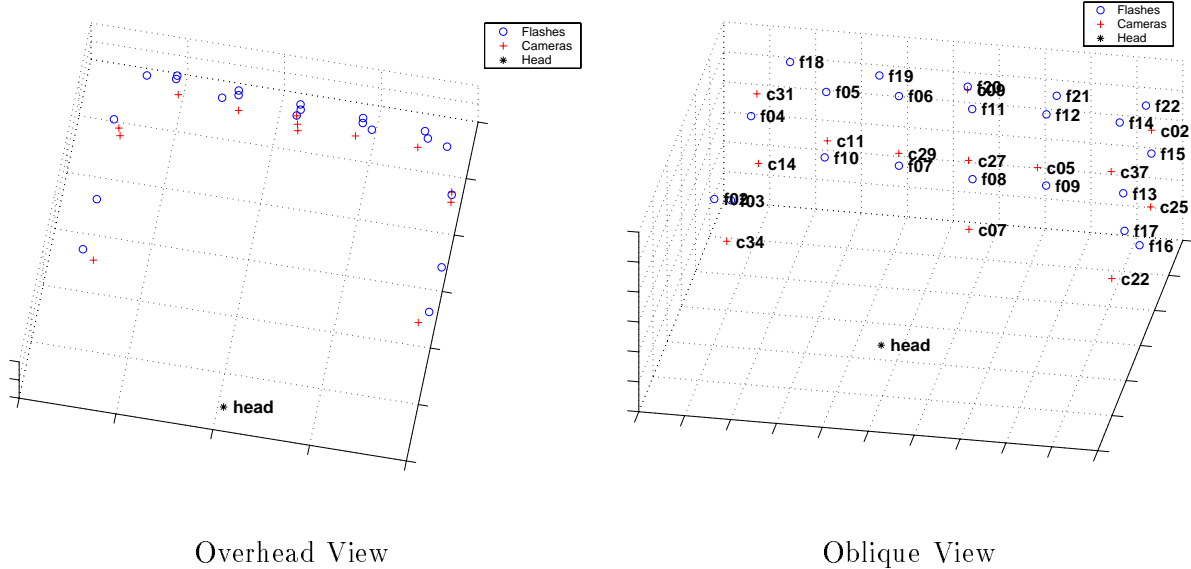


Figure 2: The xyz-locations of the head, the 13 cameras, and the 21 flashes plotted in 3D to illustrate their approximate locations. The locations were measured with a Leica theodolite.

We positioned 9 of the 13 cameras at roughly head height in an arc from approximately a full left profile to a full right profile. Each neighboring pair of these 9 cameras are therefore approximately  $22.5^\circ$  apart. Of the remaining 4 cameras, 2 were placed above and below the central (frontal) camera, and 2 were placed in the corners of the room where a typical surveillance camera would be. The locations of 10 of the cameras can be seen in Figure 1. The other 3 are symmetrically opposite the 3 right-most cameras visible in the figure. Finally, we measured the locations of the cameras using a theodolite. The measured locations are shown in Figure 2. The numerical values are included in the database.

The pose of a person’s head can only be defined relative to a fixed direction, most naturally the frontal direction. Although this fixed direction can perhaps be defined using anatomical measurements, even this method is inevitably somewhat subjective. We therefore decided to define pose by asking the person to look directly at the center camera (c27 in our numbering scheme.) The subject therefore defines what is frontal to them. In retrospect

this may have been a mistake because some of the subjects clearly introduce an up-down tilt or a left-right twist. The absolute pose measurements that can be computed from the head position, the camera position, and the frontal direction (from the head position to camera c27) should therefore not be completely trusted. The relative pose, on the other hand, can be trusted because the cameras were not moved. The PIE database can therefore be used to evaluate the performance of head pose estimation algorithms either using the absolute head poses, or using the relative poses to estimate how internally consistent the algorithms are.

Another issue that has to be dealt with when using multiple cameras is photometric (i.e. intensity and color response) calibration. We first switched off all automatic gain and gamma correction. The apertures and shutter speeds are therefore constant across the entire database. Although the cameras are of the same type, there is still a large amount of variation in their photometric responses, both due to their manufacture and due to the fact that the iris setting on the cameras were all set manually and therefore may vary slightly from camera to camera. We “auto white-balanced” the cameras, but there is also some noticeable variation in the color responses of the cameras. To allow the cameras to be intensity (gain and bias) and color calibrated, we captured images of color calibration charts and include them in the database. Although we do not know “ground-truth” for the colors, the images can be used to equalize the color (and intensity) responses across the 13 cameras.

## **2.2 The Computer Controlled Flash System: Illumination**

To obtain the illumination variation we extended the 3D Room with a “flash system” similar to the Yale Dome used to capture the data shown in [Georghiades *et al.*, 2000]. With help from Athinodoros Georghiades and Peter Belhumeur, we used an Advantech PCL-734, 32

channel digital output board to control 21 Minolta 220X flashes. The Advantech board can be directly wired into the “hot-shoe” of the flashes. Generating a pulse on one of the output channels then causes the flash to go off. We placed the Advantech board in one of the 17 computers used for image capture in the 3D Room and integrated the flash control code into the image capture routine so that the flash, the duration of which is approximately 1ms, occurs while the shutter (duration approximately 16ms) is open. We then modified the image capture code so that one flash goes off in turn for each image captured. We were then able to capture 21 images, each with a different illumination condition, in  $21/30 \approx 0.7\text{sec}$ .

In the Yale illumination database [Georghiades *et al.*, 2000] the images are captured with the room lights switched off. The images in the database therefore do not look entirely natural. Most of the time you see people with approximately neutral illumination, perhaps with one or two other sources. To obtain representative images of such cases (that are more appropriate for determining the robustness of face detection and recognition algorithms to illumination change) we decided to capture images both with the room lights on and with them off. We include images with the lights off to duplicate the Yale database and to provide images to do photometric stereo on. To get images that look reasonably natural the room illumination and the flashes need to contribute approximately the same amount of light in total. (The flash is much brighter, but is illuminated for a much shorter period of time.)

Even so, we still found it necessary to place blank pieces of paper in front of the flashes as a filter to reduce their brightness. The iris setting is then set so that without the flash the brightest pixel registers a pixel value of around 128 and that with the flash only a small number of pixels saturate. (Since the “color” of flashes is quite “hot,” it is only the blue channel that ever saturates. The database therefore contains saturated data that is

useful for evaluating the robustness of algorithms to saturation, and also the red and green channels that are unsaturated and which can be used for tasks, such as photometric stereo, that require unsaturated data.)

An extra benefit of this decision is that the flashes are then substantially less bright than when un-masked. There are therefore very few cases of the subjects either blinking or grimacing during the capture sequence. The fact that the flashes are further away from the subject than in the Yale Dome also helps in this regard. On the other hand, a slight disadvantage of this decision is that the images that were captured without the flashes are compressed into 0-128 intensity levels and so appear fairly dark. This can easily be corrected for at the cost of increased pixel noise. Since we used fairly high quality cameras, however, the signal to noise ratio should still be relatively high.

## **2.3 The Capture Procedure: Expression**

The database was captured over a period of months. For most of this time the cameras were not touched, with one exception when somebody borrowed one of the cameras. We therefore split the database into two sessions, before and after this event. (If we capture more people in future, these will form new sessions.) Within each session the cameras are kept in the same locations and we made every effort to keep the settings the same. Each session does however last a number of days and so we cannot guarantee some settings did not change.

For this reason, we captured new color calibration images every day. Although this is probably unnecessary, the color and iris settings on the cameras are liable to drift, or to be accidentally changed. If you decide to color calibrate the cameras, it is probably fine to assume that the settings are the same for each session. If you want to be really sure,

you should calibrate using the images captured on the same day. We also captured a set of background images every day. (We captured the background images since they may help in determining the position of the head in certain evaluation tasks.) Again, use the background images captured on the same day if you want the best performance.

To obtain the expression (and illumination) variation, we asked each of the subjects to perform several tasks. After the subjects signed the various paperwork (i.e. the consent form and gave us contact information), we then led them through the following steps:

**Neutral:** We asked the person to sit in the chair and look at the central camera with a neutral expression. We then captured a single image from each of the 13 cameras. In this step, the room lights are left on and there is no illumination variation.

**Smile:** We repeated the previous step, but asked the subject to smile.

**Blink:** We again repeated the previous steps, but asked the subject to close their eyes to simulate a blink. Again all 13 cameras were used.

**Without Glasses:** If the subject wears glasses, we now asked them to remove them. We then captured a single image from all 13 cameras with them in a neutral expression.

**With Room Lights:** Next we captured the illumination variation with the room lights switched on. We asked the person to sit in the chair with a neutral expression and look at the central (frontal) camera. We then captured 24 images from each camera, 2 with no flashes, 21 with one of the flashes firing, and then a final image with no flashes. If the person wears glasses, we got them to keep them on. (Although we captured this data from each camera, for reasons of storage space we decided to only keep the output of three cameras, the frontal camera, a 3/4 profile, and a full profile view.)

**Without Room Lights:** We repeated the previous step but with the room lights off. Since these images are likely to be used for photometric stereo, we asked the person to remove their glasses if they wear them. (We kept the images from all cameras this time.)

**Talking:** We asked the person to look at the frontal camera and start counting, “1, 2, 3, etc.” We then captured 2 seconds (60 frames) from 3 cameras (frontal, 3/4 profile, and full profile) of them talking. The person wears their glasses if they have them. As in the first 4 step, the room lights are left on and the flashes are not fired.

On average this procedure took about 10 minutes for each subject. In that time, we captured (and retained) over 600 images from 13 poses, with 43 different illuminations, and with 4 expressions. The total data storage requirement is approximately 600MB per person in color “raw .ppm” images. The images are  $640 \times 486$  color images. (The first 6 rows of the images contain synchronization information added by the VITC units in the 3D Room [Kanade *et al.*, 1998]. This information could be discarded.) The total storage requirement for 68 people is therefore around 40GB. This can of course be reduced by compressing the images.

### 3 Database Organization, Example Images, and Uses

The CMU PIE database is organized as a collection of images for each subject. The data is then organized into expression images, talking images, illumination variation with the room lights on, illumination variation with the room lights off, miscellaneous calibration and background images, and personal attributes. All images have pose variation. We now provide some example images and describe some possible uses of the data.

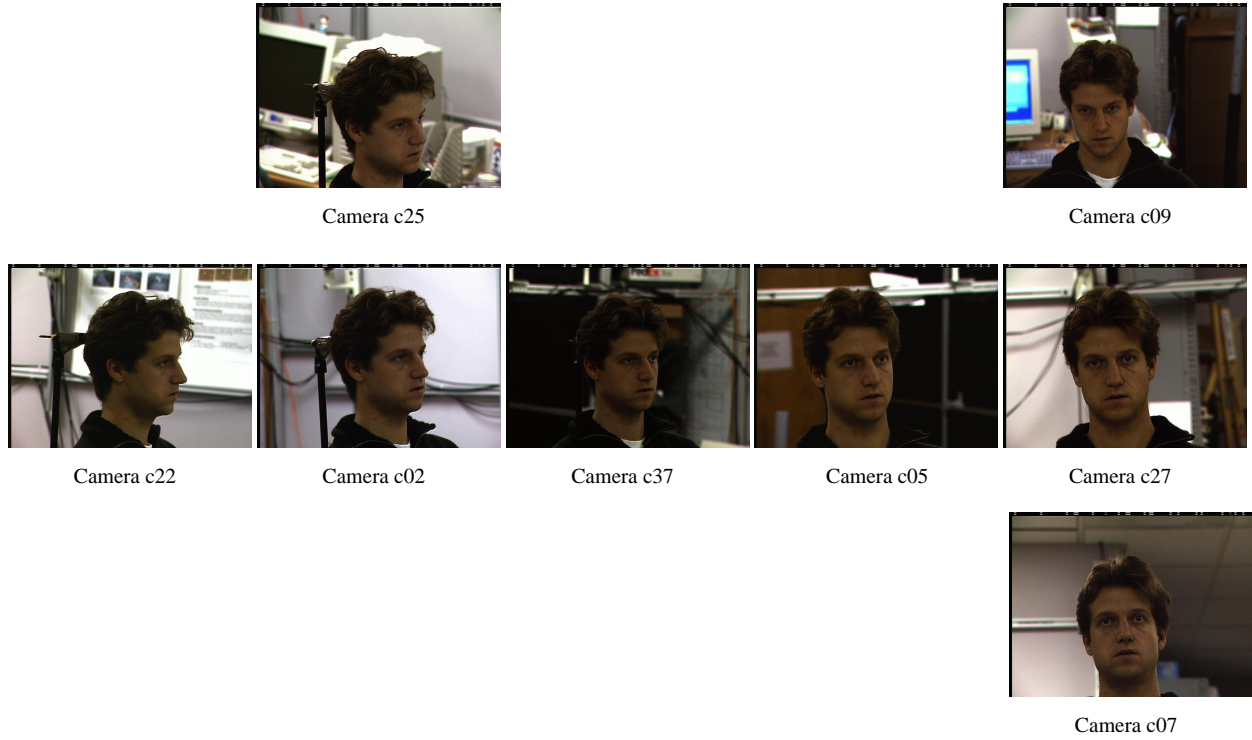


Figure 3: An example of the pose variation in the CMU PIE database. We show images of one person from 8 of the 13 cameras. The other 5 cameras are arranged symmetrically to the 5 cameras on the left. (See also Figure 2 for the camera locations.) As can be seen, the pose varies from full left profile to full frontal and on to full right profile (not shown.) The 9 cameras in the horizontal sweep are each separated by approximately  $22.5^\circ$ . The 4 other cameras include 2 above and below the central camera, and 2 in the corners of the room, a typical location for surveillance cameras.

### 3.1 Pose and Expression Variation

An example of the pose variation in the PIE database is shown in Figure 3. This figure contains an image of one subject in the database from 8 of the 13 cameras. The other 5 cameras are arranged symmetrically. As can be seen, there is a wide variation in pose from full profile to full frontal. This subset of the data should be useful for evaluating the robustness of face recognition algorithms across pose. Since the camera locations are known, it can also be used for the evaluation of pose estimation algorithms. Finally, it might be useful for the evaluation of face recognition algorithms that combine information from

multiple widely separated views to enhance their recognition performance. An example of such an algorithm would be one that combines frontal and profile views to do face recognition.

An example of the expression variation is shown in Figure 4. The subject is asked to provide a neutral expression, to smile, to blink (i.e. they are asked to keep there eyes shut for a while), and to talk. For neutral, smiling, and blinking, we keep 13 images, one from each camera. For talking, we keep 2 seconds of video (60 frames.) Since this occupies a lot more space, we only keep the data for 3 cameras, frontal camera c27, 3/4 profile camera c22, and full profile camera c05. In addition, for subjects who usually wear glasses, we collect one extra set of 13 images of them across pose without them wearing their glasses. (Here the subject is asked to put on a neutral expression.)

The expression variation data is designed to test the robustness of face recognition algorithms to expression (and pose.) This is the reason we chose to use expressions that appear most frequently in the real world. Most of the time people either have a neutral expression, or are smiling, blinking, or talking. Showing disgust, anger, etc, occur far less frequently. These expressions are also covered in other databases such as [Kanade *et al.*, 2000]. A special reason for including blinking was because many face recognition algorithms use the eye pupils to align a face model to. It is therefore possible that they are particularly sensitive to subjects blinking. We can now test whether this is indeed the case.

## 3.2 Illumination Variation

Examples of the illumination variation are shown in Figures 5 and 6. Figure 5 contains the variation with the room lights on and Figure 6 with the lights off. Comparing the images, we see that those in Figure 5 appear more natural and representative of images that appear




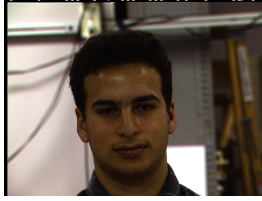
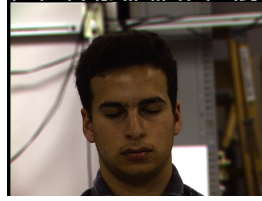
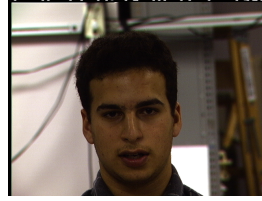


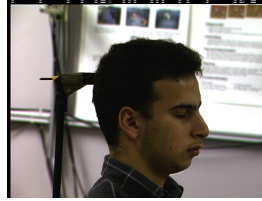
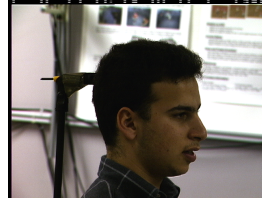




c27				
c22				
c05				
	Neutral	Smiling	Blinking	Talking

Figure 4: An example of the expression variation in the PIE database. Each subject is asked to give a neutral expression, to smile, to blink, and to talk. We capture this variation across pose. For the neutral images, the smiling images, and the blinking images, we keep the data for all 13 cameras. For the talking images, we only keep 2 seconds (60 frames) of video from each of three cameras (frontal c27, 3/4 profile c22, and full profile c05). In addition, for subjects who wear glasses we also capture one set of 13 neutral images of them without their glasses. (Not shown.)

in the real world. On the other hand, the data with the lights off was captured to reproduce the Yale database [Georghiades *et al.*, 2000]. This will allow a more direct comparison to be made between the two databases than otherwise would be possible. Besides the room lights, the other major differences between these two parts of the database are: (1) the subjects wear their glasses in Figure 5 (if they have them) and not in Figure 6, and (2) in Figure 6 we retain all of the images, whereas for Figure 5 we only keep the data from 3 cameras, the frontal camera c27, the 3/4 profile camera c22, and the full profile camera c05. We foresee a number of possible uses for the illumination variation data. First it can be used to reproduce










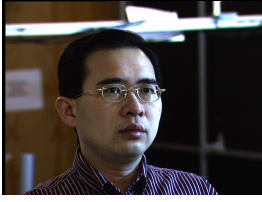


c27				
c22				
c05				
	Room Lights	Flash f01	Flash f09	Flash f17

Figure 5: An example of the illumination variation with the room lights on. The subject is asked to pose with a neutral expression and to look at the central camera. We then capture 24 images of them (for each camera), 2 with just the background illumination, 21 with one of the flashes firing, and one final image with just the background illumination. Notice how the combination of the background illumination and the flashes leads to much more natural looking images than with just the flash. (See Figure 6 for comparison.) For lack of space, we only kept the data from 3 cameras.

the results in [Georghiades *et al.*, 2000]. Secondly it can be used to evaluate the robustness of face recognition algorithms to pose and illumination in a systematic manner.

A natural question that arises is whether the data with the room lights can be converted into that without the lights by simply subtracting one of the images with no flashes and just the background illumination from one of the images with both. Preliminary results indicate that this is the case. For example, Figure 7 contains an image with just the room lights and another image taken with both the room lights and one of the flashes a short fraction of a second later. We also show the difference between these two images and compare

c27				
c22				
c05				
	Flash f01	Flash f09	Flash f17	Flash f21

Figure 6: An example of the illumination variation with the room lights off. This part of the dataset largely corresponds to the Yale illumination database [Georghiades *et al.*, 2000]. We captured it to reproduce their results, and to allow direct comparison between the two databases. This part of the dataset is perhaps less representative of facial images that appear in the real world than that captured with the room lights off but can be directly used for photometric stereo.



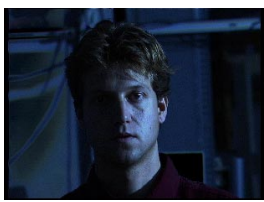

			
Room Lights	With Flash	Difference	Flash Only

Figure 7: An example of taking an image with the room lights on and a single flash and subtracting an image with just the room lights on taken a short fraction of a second earlier. The difference image is compared with an image taken with the same flash and with the room lights off. Although the facial expression is a little different, otherwise the images appear similar. (There are a few noticeable differences in the background caused by the images saturating there.)

it with an image of the same person taken with just the flash; i.e. with the room lights off. Except for the fact that the person has a slightly different expression (that image was taken a few minutes later), the images otherwise look fairly similar. We have yet to try to see whether other algorithms behave similarly on these two images. If this works, we will then investigate forming synthetic images of a person captured under multiple flashes and adding them to the database as additional images to evaluate algorithms on.

### **3.3 Personal Attributes, Calibration Data, and Backgrounds**

Besides the main pose, illumination, and expression variation data we also collected a certain amount of other images to aid in calibration and other processing. For example, at the beginning of each day we captured a background image and a set of color calibration images. An example of a background image is shown in Figure 8. The background image is then subtracted from an image taken on that day. As can be seen, background subtraction works very well. The background images may well therefore be helpful in locating the head during certain empirical evaluations. An example color calibration image is also shown in Figure 8. Although we do not have ground-truth for the colors on the chart, the same chart is used in all of the images. The color calibration image can therefore be used to estimate a simple gain and bias model for each color channel separately to give a first order color calibration. Finally, we also include some information about the subjects in the database. For each subject we record their sex, their age, whether they wear glasses, whether they have a mustache, whether they have a beard, and the date on which they were imaged. At the time of writing, we have not decided whether or not to include the “race” or “ethnicity” of the subjects in the personal attributes.

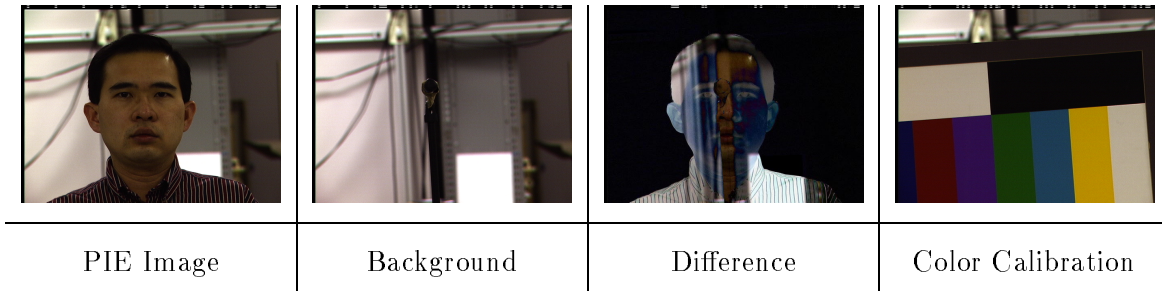


Figure 8: An example of a background image and a demonstration of how background subtraction can be used to locate the face. This may be useful in certain evaluations to locate the face. We also include an example of a color calibration image. These images can be used to estimate simple linear response functions for each of the 3 color channels independently.

## 4 Obtaining the PIE Database

Information on how to obtain the database can be obtained from the PIE Database webpage. Please go to the Robotics Institute webpage at URL [www.ri.cmu.edu](http://www.ri.cmu.edu) and search for “PIE Database.”

## Acknowledgements

We would like to thank Athinodoros Georgiades and Peter Belhumeur for providing us with the details of how they constructed the Yale “flash system.” We largely followed their design. Sundar Vedula and German Cheung gave us great help using the CMU 3D Room. We would also like to thank Henry Schneiderman and Jeff Cohn for discussions on what data to collect. Financial support for the extension of the 3D Room and the collection of this database was provided by the U.S. Office of Naval Research (ONR) under contract N00014-00-1-0915.

## References

- [Georghiades *et al.*, 2000] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 277–284, Grenoble, France, March 2000.
- [Kanade *et al.*, 1998] Takeo Kanade, Hideo Saito, and Sundar Vedula. The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams. Technical Report CMU-RI-TR-98-34, Robotics Institute, Carnegie Mellon University, December 1998.
- [Kanade *et al.*, 2000] Takeo Kanade, Jeffrey Cohn, and Ying-Li Tian. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.
- [Philips *et al.*, 1997] P.J. Philips, H. Moon, P. Rauss, and S.A. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, pages 137–143, San Juan, Puerto Rico, 1997.