A Video Database of Moving Faces and People

Alice J. O'Toole, Joshua Harms,
Sarah L. Snow, Dawn R. Hurst,
Matthew R. Pappas,
Janet H. Ayyad, and
Hervé Abdi

Abstract—We describe a database of static images and video clips of human faces and people that is useful for testing algorithms for face and person recognition, head/eye tracking, and computer graphics modeling of natural human motions. For each person there are nine static "facial mug shots" and a series of video streams. The videos include a "moving facial mug shot," a facial speech clip, one or more dynamic facial expression clips, two gait videos, and a conversation video taken at a moderate distance from the camera. Complete data sets are available for 284 subjects and duplicate data sets, taken subsequent to the original set, are available for 229 subjects.

Index Terms—Face database, face recognition, face tracking, digital video.

1 Introduction

Research on computer-based face recognition has been an active area of study in recent years. This research has applications in person identification and verification for security systems [1], [2], facial expression analysis, e.g., [3], [4], [5], [6], [7], [8], and face classification, e.g., [9]. To date, most of this research relies on static images of faces and people, taken under relatively controlled viewpoint and illumination conditions. Further, recognition matches are generally made between similar kinds of images, usually facial snapshots taken in close temporal proximity so that changes in surface aspects of appearance (e.g., hairstyle) are minimal. Within these constraints, enormous progress has been made in the past decade in solving the complex problems involved in face recognition [1], [2].

The most useful current and future applications of face and person recognition research, however, lie in more naturalistic contexts. For these applications, algorithms are required to operate on data from video cameras that capture the natural movements of people, often in public places [10], [11]. In most of these settings, the illumination comes from a natural source such as sunlight, for which the direction and strength of the light varies by time of day, season of the year, and weather conditions. This problem of recognition in natural illumination has been identified as an important area of future research [12]. A further complication is that target images must be matched to images from a database that may have been taken months or even years previously.

One reason for the relatively limited amount of research on recognition in naturalistic settings is the limited availability of image/video databases that provide large numbers of people in a diversity of contexts and imaging situations though see [13], [14]. Such databases are needed to test the accuracy of computational models of face and person recognition when the people (and their faces) are in motion and when the learning and test images/videos are of different types. For example, one common application might be to match a frontal mug shot of a face to a person walking past a surveillance camera in a public place. Another common application is to match videos of people walking in different directions with respect to the camera. Motion tracking of the person, head, face, mouth/lips are also important prerequisite steps for many recognition applications.

In addition to recognition-based applications, there are applications in human-computer interface design and computer graphics for a database containing natural face and body movements, e.g., [15], [16], [17], [18], [19], [20], [21]. Facial expressions and gestures provide humans with information relevant for social interaction and intent. This information is potentially useful in designing interactive computer systems that adapt their responses to the needs and desires of a user. Such applications are useful at a person-based level as well. Being able to categorize the movements of a person walking as "deliberate," "aimless," "rushed," or "fatigued," can be helpful in certain applications.

In this paper, we describe a database of image and video clips of faces and people. The database was developed for testing the effects of motion on human memory for faces and people, but it is also useful for testing automatic systems. In fact, evaluating automatic face and person recognition systems is best done with accurate information about human performance on comparable tasks, e.g., [22]. Parts of this database have been used recently in the Face Recognition Vendor Test 2002 of 10 commercial and mature prototype face recognition systems [12]. Publicly available databases can provide a standard by which the accuracy of algorithms and human observers can be assessed and compared.

2 Database Definition

The database consists of static digital images and video clips of faces and people. The static images and facial videos were taken at close range, under controlled lighting conditions in an indoor laboratory environment. The video clips of people walking and conversing were taken under variable illumination conditions and at moderate and varying distances. Specifically, these videos were taken in a building foyer with high ceilings, enclosed entirely on one side with glass windows. This environment approximates outdoor lighting conditions, while protecting the subjects and the cameras from the elements. A full session of data includes still and video images of an individual, as described in the sections below. A duplicate session includes a full set of these still and video images, taken between one week and six months subsequent to the original set. More precisely, the average interval between the first and second sessions was 24.1 days and the median interval was 7.0 days. More precise filming details, such as camera distance, etc., are given in the Appendix. Still and video color examples of the data types can be viewed at our Web site.[1]

2.1 Still Images

The *facial mug shots* are high quality static images that approximate the mug shot style images available in many face databases. The mug shots provide nine discrete views of the face, ranging from a

• *The authors are with the School of Behavioral and Brain Sciences, GR4.1, University of Texas at Dallas, Richardson, TX 75083-0688. E-mail: {otoole, herve}@utdallas.edu.*

Manuscript received 4 Dec. 2003; revised 17 Sept. 2004; accepted 4 Nov. 2004; published online 11 Mar. 2005.
Recommended for acceptance by R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0396-1203.

1. www.utdallas.edu/dept/bbs/FACULTY_PAGES/otoole/database.htm.

Fig. 1. Row 1 shows a *facial mug shot* series with nine still images, varying in pose from left (- 90 degrees) to right (90 degrees) profile in 22.5-degree steps. The second row contains five still images extracted from a *facial speech* video. The third and fourth rows contain images extracted from a *disgust expression* and *laughter expression* video, respectively.

left to right profile, in equal-degree steps (Fig. 1). To assure comparable views for each subject, numbered markers were suspended from the ceiling at the appropriate angles to be used as fixation points by the subject being filmed. Additionally, each participant wore a gray smock that covered any clothing that was visible to the camera.

2.2 Videos of Faces at Close Range

The *dynamic facial mug shots* provide a moving version of the facial mugs shots just described. In these videos, the head moves in a natural way through the same nine viewpoints used for the mug shots. Most subjects appear to be systematically scanning a room (left to right) for someone or something. These videos were taken at the same distance as the mug shots. The subject was instructed to turn their head pausing briefly at each of the nine angles used for the mug shot images. These clips are 10 seconds in length. To assure comparable timing for the models, we used a metronome set at 1-second intervals to cue the subject's movement to the next fixation marker.

The *dynamic facial speech* videos capture the rigid and nonrigid movements we make when we speak. Most faces in the database include both a "neutral" and an "animated" facial speech clip (Fig. 1). Animated clips include one or more head motions (tilts, etc.), facial expressions and eye gaze changes in addition to the speech movements. In each instance, the subject was filmed while responding to a series of mundane questions. The sound was removed from these clips, so the files contain videos of the subjects speaking without an accompanying sound track. These clips are 10 seconds in length.

The *dynamic facial expression* clips capture emotions such as happiness, sadness, and disgust. These are common nonrigid movements of the face. We employed a simple method to capture dynamic, natural facial expressions. During filming, the subject watched a 10-minute video, which contained scenes from various movies and television programs intended to elicit different emotions. The digital stream captured during the 10-minute filming session was scanned subsequently for instances of nonrigid facial motions that corresponded (by the judgment of the experimenter coding the data[2]) to: happiness, sadness, fear, disgust, anger, puzzlement, laughter, surprise, boredom, or disbelief (Fig. 1). It is important to note that the expression rating was not done formally or by rigorous experimental procedures. Indeed, without making additional assumptions about how to determine what constitutes a "smile" or "disgust" expression (e.g., [23]), there can be no ground truth for the expression videos. Thus, researchers are advised to carry out psychological expression-norming procedures prior to making claims about particular facial expressions found in the database.

On average, three expressions were captured for each individual in each session. The expression segments were edited into 5-second video clips. This was difficult because expressions varied in length. Some occurred over a few frames, others lasted many seconds, and some spanned the full 5-second standardized clip time. In cases where the expression duration was shorter than the clip duration, we centered the expression to the middle of the clip. In other cases, the clip begins and ends with an expressing face.

We also captured a 5-second "blank stare" video, containing no explicit facial motions, but other natural movements of the head and eye blinks. Fig. 2 shows the number of instances of various expressions in the database.

The expression clips differ from previously available facial expression databases in several ways. First, as noted, the facial expressions have not been verified or normed in a formal sense as being instances of one the primary expressions defined by Ekman and Friesen [23]. Second, most of the expressions are more subtle than those available previously, though see Pantic and Rohkrantz [24] for a review of the range of face images used in automatic analysis of facial expression. Third, because these are dynamic stimuli, head and eye movements often accompany the expressions. Finally, some clips contain more than one expression (e.g., a puzzled expression, which turns to surprise or disbelief, and ultimately laughter). In these cases, we adapted file-naming conventions to indicate the presence of multiple expressions in a clip.

Combined, the close-range videos provide test stimuli for face recognition and tracking algorithms that operate when the head is undergoing rigid and/or nonrigid transformations. The dynamic mug shots, speech, and expression videos are likewise useful for computer graphics modeling of heads and facial animation.

2. Multiple experimenters coded different data sets, though only one experimenter coded each individual video clip.

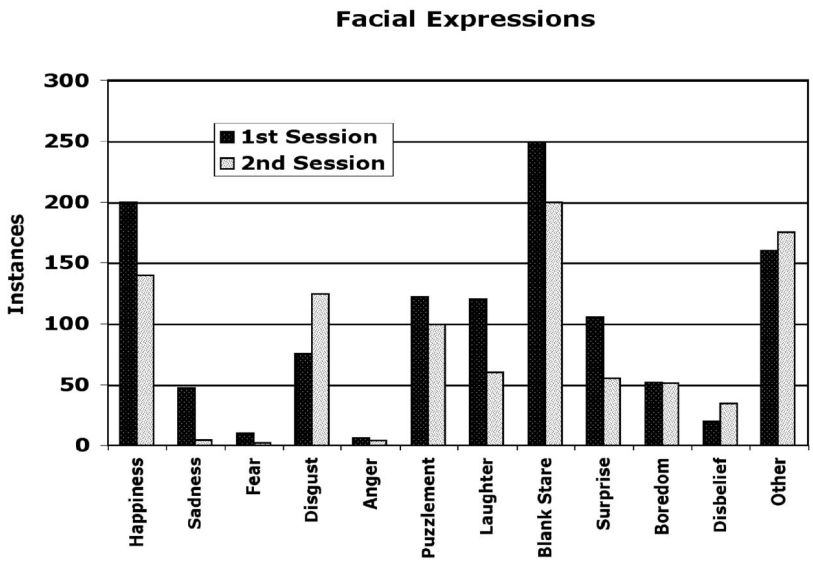

Fig. 2. Facial expressions in the database. There were 284 subjects in the first session and 229 subjects returned for a second session. "Other" refers to expressions not catalogued in the database, or duplicate examples of expressions already catalogued.

2.3 Videos of People at Moderate Distances

In the *parallel gait* video, the subject walks parallel to the line of sight of the camera, approaching the camera, but veering off to the left in the final few paces (Fig. 3). These videos capture the subject from the start point until he/she passes out of view. Thus, the time varies somewhat for each model, but lasts approximately 10 seconds for most subjects. The *perpendicular gait video* captures the subject walking perpendicular to the line of sight of the camera at a distance (see the Appendix for details). The video begins with the subject walking out from behind a wall partition and lasts until the subject passes out of the camera view (4-6 seconds) behind a second wall partition (Fig. 3).

The *conversation video* shows a conversation between the subject and a laboratory staff member. The lab member stands with his/her back to the camera and the subject faces the lab member. The camera is placed at a moderate distance from the pair and slightly overhead (Fig. 3). To capture some natural gesturing in these videos, the subject was asked to give directions to a building on campus. These videos last 10 seconds.

3 Demographics of the Database

Students from The University of Texas at Dallas participated as subjects. More female than male students (Males = 76, Females = 208) volunteered. Most subjects were Caucasians, between the ages of 18 and 25. Participants described as "other" did not belong to an ethnic group represented in our database. Subject's ethnicity was defined by an optional questionnaire response. The ethnicity and gender distributions of the subjects in the two sessions were comparable (Fig. 4).

4 Summary

The database contains a variety of still images and videos of a large number of individuals taken in a variety of contexts. A second duplicate session is available for most subjects, allowing for recognition tests that make use of images and videos in which the subject may have a different hairstyle, different clothing, and may be otherwise different in appearance. This database is useful for testing the performance of humans and machines on the tasks of face/person recognition, tracking, and computer graphics modeling of natural human motions.

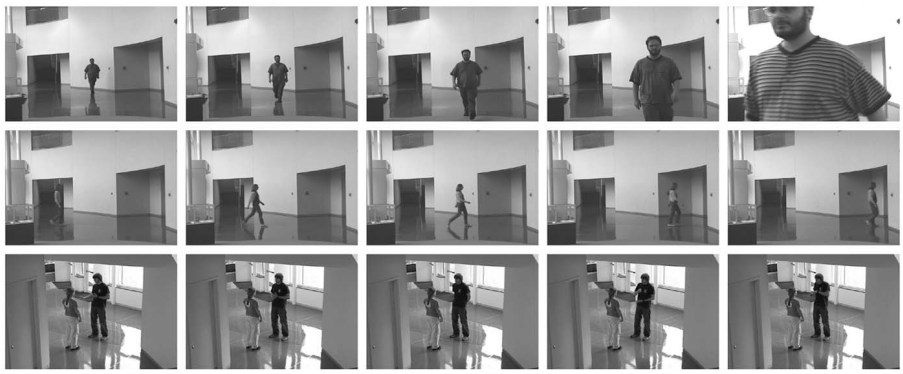

Fig. 3. The first row of the figure contains five still images extracted from a *parallel gait* video. The second row contains five still images extracted from a *perpendicular gait* video. The third row of the figure contains five still images extracted from a *conversation* video.

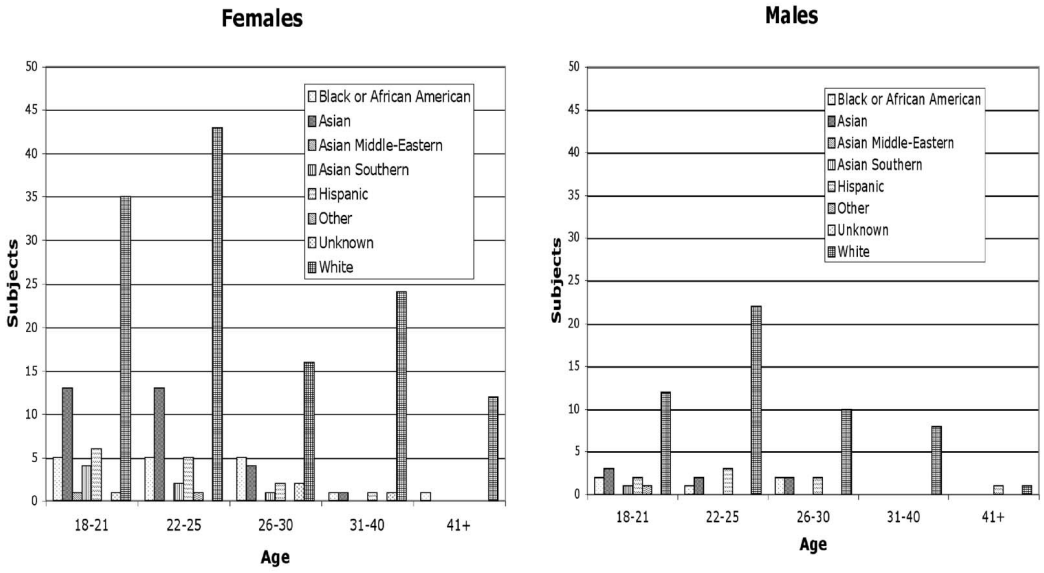

Fig. 4. Left: demographics of 208 female subjects. Right: demographics of 76 male subjects.

TABLE 1
Storage Breakdown by Image Type and Duration

Image/Video Type	Storage	Duration
Mug shots	3.6 mb	NA
Exploration	40.1 mb	10 s.
Speech	40.1 mb	10 s.
Facial Expressions	21.9 mb	5 s.
Parallel gait	32.0-44.0mb	10 s.
Perpendicular gait	18.2 mb	4-6 s.
Conversation	40.1 mb	10 s.

APPENDIX

Equipment. The images and videos were collected using a Canon Optura Pi digital video camera.[3] The Optura Pi employs a single progressive scan CCD digitizer that produces minimal motion-aliasing artifacts.

Close range face videos. The camera was placed at a distance of 2 meters, directly in front of the participant. The illumination approximated ambient lighting. Specifically, the photographic set-up consisted of three[4] 500 watt, 12 inch flood lights mounted on stands at a height of 186 centimeters and set apart from each other by 1.7 meters. Each light was 2.3 meters from the participant. Three EBW no. 2 blue corrective lamps were used with the floods to avoid the reddish tendencies of standard tungsten lighting. A clip-on diffusion screen was mounted on the front of each flood to soften the light. In addition, a neutral gray background paper was mounted on a wall behind the participant. Each participant wore a gray smock that covered any clothing that was visible to the camera.

Moderate range gait and conversation videos. These videos were taken in a foyer with large panel windows that allowed for natural variations in illumination. Camera placement in the parallel videos was 13.6 meters from the start point to the camera. Thus, the subject's distance from the camera varied from 13.6 meters to approximately 1 or 2 meters, at the point where the subject veered off to the left of the camera. For the perpendicular videos, the distance between the camera and the center point of the subject's trajectory was 10.4 meters. The conversation videos were filmed from the top of a short flight of stairs at a height of 3.5 meters, looking down on the subject and lab member. The distance of the subject to the camera was approximately 8 meters.

3. When the project began, a Canon XL-1 digital video camera was used to capture the close-range face images, while the Canon Optura Pi was used for the gait and conversation images. After filming approximately 62 participants, we elected to switch over to the Optura Pi for all image capturing because of the Optura Pi's progressive scanning capability and easier portability. As a consequence, there is a difference in overall image aesthetic between our earlier images and those captured after the changeover. Files are not explicitly marked with the camera used to create them.

4. Because of technical problems during the collection process, only two of the floodlights were used on certain occasions. The difference in appearance is minimal and we estimate that such images constitute less than 5 percent of all images in the database.

File Format. Still images were exported from Final Cut Pro (FCP) to TIFF format at a resolution of 720 by 480 with 32-bit color. The videos are stored in DV Stream format at the same resolution but with 24-bit color and 29.97 frames per second (see Table 1).

AVAILABILITY

The database is available from the authors. We maintain a searchable database in Microsoft Access that will be made available upon request. We will provide a brief key explaining the file naming conventions used with the various file types. This database is for noncommercial use only, as the consent forms signed by the subjects allow use only for research. A small number of subjects have additionally granted permission for their faces to appear in research publications. Requesters of the database will be required to sign a form agreeing to the terms of use and to respecting the limits of the subjects' consent. Given the size of the database, the requester will be required to supply a 160-gigabyte hard disk and will be responsible for handling and postage.

ACKNOWLEDGMENTS

This work was supported by a grant from the Human ID Project (DARPA/DOD) to A.O'T and H.A.

REFERENCES

[1] P.J. Phillips, H Moon, S. Rizvi, and P. Rauss, "The Feret Evaluation Methodology for Face Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 1090-1103, 2000.

[2] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face Recognition: A Literature Review," Technical Report CAS-TR-948, Univ. of Maryland, College Park, Oct. 2000.

[3] J.F. Cohn, A.J. Zlochower, J. Lien, and T. Kanade, "Automated Face Analysis by Feature Point Tracking has High Concurrent Validity with Manual FACS Coding," *Psychophysiology,* vol. 36, pp. 35-43, 1999.

[4] S. Dubuisson, F. Davoine, and M.A. Masson, "Solution for Facial Expression Representation and Recognition," *Signal Processing-Image Comm.,* vol. 17, pp. 657-673, 2002.

[5] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition,* vol. 36, pp. 259-275, 2003.

[6] J.J. Lien, T. Kanade, J.F. Cohn, and L. Ching-Chung, "Subtly Different Facial Expression Recognition and Expression Intensity Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 853-859, June 1998.

[7] J.J. Lien, T. Kanade, J.F. Cohn, and L. Ching-Chung, "Automated Facial Expression Recognition Based on FACS Action Units," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 390-395, Apr. 1998.

[8] Y. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 97-115, Feb. 2001.

[9] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Toward Automatic Simulation of Aging Effects on Face Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, pp. 442-455, 2002.

[10] R.T. Collins, A.J. Lipton, and T. Kanade, "Introduction to the Special Section on Video Surveillance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 745-746, Aug. 2000.

[11] L.A. Wang, W.M. Hu, and T.N. Tan, "Recent Developments in Human Motion Analysis," *Pattern Recognition,* vol. 36, pp. 585-601, 2003.

[12] P.J. Phillips, P. Grother, R. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report," NISTIR 6965, www.frvt.org, 2003.

[13] T. Kanade, J. Cohn, and Y.-L. Tian, "Comprehensive Database for Facial Expresison Analysis," *Proc. Fourth IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2000.

[14] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, pp. 1615-1618, 2003.

[15] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," *Proc. SIGGRAPH '99,* 1999.

[16] Q. Chen and G. Medioni, "Building 3-D Human Face Models from Two Photographs," *J. Visual Signal Processing Systems for Signal Image and Video Technology,* vol. 27, pp. 127-140, 2001.

[17] T.F. Cootes, G.F. Edwards, and C.J. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, pp. 681-685, 2001.

[18] P.Y. Hong, Z. Wen, and T.S. Huang, "Real-Time Speech-Driven Face Animation with Expressions Using Neural Networks," *IEEE Trans. Neural Networks,* vol. 13, pp. 916-927, 2002.

[19] F. Pighin, R. Szeliski, and D.H. Salesin, "Modeling and Animating Realistic Faces from Images," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 143-169, 2002.

[20] K.K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, pp. 39-51, 1998.

[21] C.Z. Zhang and F.S. Cohen, "3-D Face Structure Extraction and Recognition from Images Using 3-D Morphing and Distance Mapping," *IEEE Trans. Image Processing,* vol. 11, pp. 1249-1259, 2002.

[22] S.M. Snow, G.J. Lannen, A.J. O'Toole, and H. Abdi, "Memory for Moving Faces: Effects of Rigid and Non-Rigid Motion," *J. Vision,* vol. 2, no. 7, p. 600a, 2002.

[23] P.J. Ekman and W.V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco: Consulting Psychology Press, 1978.

[24] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 1424-1445, 2000.