Face Databases and Evaluation*

Dmitry O. Gorodnichy

Laboratory and Scientific Services Directorate, Canada Border Services Agency, 79 Bentley Ave., Ottawa, ON, K1A 0L5, Canada
Email: dg@videorecognition.com

Synonyms

Face Datasets; Face Recognition Performance Evaluation

Definition

Face Databases are imagery data that are used for testing face processing algorithms. In the contents of biometrics, face databases are collected and used to evaluate the performance of face recognition biometric systems.

Face recognition evaluation is the procedure that is used to access the recognition quality of a face recognition system. It involves testing the system on a set of face databases and/or in a specific setup for the purpose of obtaining measurable statistics that can be used to compare systems to one another.

Introduction: factors affecting face recognition performance

While for humans recognizing a face in a photograph or in video is natural and easy, computerized face recognition is very challenging. In fact, automated recognition of faces is known to be more difficult than recognition of other imagery data such as iris, vein, or fingerprint images due to the fact that the human face is a *non-rigid* 3D object which can be *observed at different angles* and which may also be *partially occluded*. Specifically, face recognition systems have to be evaluated with respect to the following factors [19]:

1. face image resolution – face images can be captured at different resolutions: face images scanned from documents may have very high resolution, while face images captured with a video camera will mostly be of very low resolution,
2. facial image quality – face images can be blurred due to motion, lack of focus, and of low contrast due to insufficient camera exposure or aperture, especially when captured in uncontrolled environment,
3. head orientation – unless a person is forced to face the camera and look straight into it, s/he will unlikely be seen under the same orientation on the captured image,
4. facial expression – unless a person is quiet and motionless, the human face constantly exhibits a variety of facial expressions,
5. lighting conditions – depending on the location of the source of light with respect to the camera and the captured face, facial image will be seen with different illumination pattern overlaid on top of the image of the face,
6. occlusion – image of the face may be occluded by hair, eye-glasses and cloth such as scarf or handkerchief,
7. aging and facial surgery – compared to fingerprint or iris, a person face changes much more rapidly with time, it can also be changed as a a result of a make-up or surgery.

* Invited contribution to ”Encyclopedia of Biometrics” (Ed. Stan Z. Li), Springler Press, In press, 2009 (Online at http://www.videorecognition.com/doc).

There are over thirty publicly available face databases. In addition, there are Face Recognition Vendor Test (FRVT) databases that are used for independent evaluation of Face Recognition Biometric Systems (FRBS). Table 1 summarizes the features of the most frequently used still image facial databases, as pertaining to the performance factors listed above. More details about each database can be found at [23, 1, 2] and below is presented their summary. Video-based facial databases references can be found in [22].

Public Databases

One of the first and most used databases is AT&T (formerly ”Olivetti ORL”) database [3] that contains 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, with varying the lighting conditions, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All images were taken against dark homogeneous background with the subjects in an upright, frontal position.

The other most frequently used dataset is developed for FERET program [4]. A set of images was collected in a semi-controlled environment. To maintain a degree of consistency throughout the database, the same physical setup was used in each photography session. A duplicate set of images of persons already in the database was taken on a different day. For some individuals, over two years had elapsed between their first and last sittings, with some subjects being photographed multiple times.

The Yale Face Database [5] contains images of different facial expression (normal, happy, sad, sleepy, surprised, winking) and configuration (with/without glasses, light source at left / right). The Yale Face Database B provides single light source images of 10 subjects each seen under 576 viewing conditions (9 poses x 64 illumination conditions). For every subject in a particular pose, an image with ambient (background) illumination was also captured.

The BANCA multi-modal database was collected as part of the European BANCA project, which is aimed at developing a secure system with enhanced identification, authentication, and access control schemes for applications over the Internet [6]. The database was designed to test multimodal identity verification with various acquisition devices (high and low quality cameras and microphones) and under several scenarios (controlled, degraded, and adverse).

To investigate the role of face changes over time on face recognition performance, a large database is collected at the University of Notre Dame [7]. In addition to the studio recordings, two images with unstructured lighting are obtained.

For the same purpose, the University of Texas provides a large database of static digital images and video clips of faces [8] [27]. Data were collected in four different categories: still facial mug shots, dynamic facial mug shots, dynamic facial speech and dynamic facial expression. For the still facial mug shots, nine views of the subject, ranging from left to right profile in equal-degree steps were recorded. The sequence length is cropped to be 10 seconds.

The AR Face Database [9] is one of the largest datasets showing faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf).

XM2VTS Multimodal Face Database provides 5 shots for each person [10]. These shots were taken at one week intervals with drastic face changes occurring between the sessions. During each shot, people have been asked to count from ’0’ to ’9’ in their native language (most in French), rotate the head from 0 to -90 degrees, again to 0, then to +90 and back to 0 degrees. Also, they have been asked to rotate the head once again without glasses if they wear any.

CMU PIE Database is one of the largest datasets developed to investigate the affect of Pose, Illumination and Expression. It contains images of 68 people, each under 13 different poses, 43 different illumination conditions, and with 4 different expressions [11].

The Korean Face Database (KFDB) contains facial imagery of a large number of Korean subjects collected under carefully controlled conditions [12] . Similar to the CMU PIE database, this database has images with varying pose, illumination, and facial expressions. In total, 52 images were obtained per subject. The database also contains extensive ground truth information. The location of 26 feature points (if visible) is available for each face image.

CAS-PEAL Face Database is another large-scale Chinese face database with different sources of variations, especially Pose, Expression, Accessories, and Lighting [13].

FRVT Databases

Face Recognition Vendor Tests (FRVT) provide independent government evaluations of commercially available and prototype face recognition systems [2]. These evaluations are designed to provide U.S. government and law enforcement agencies with information to assist them in determining where and how facial recognition technology can be best deployed. In addition, FRVT results serve to identify future research directions for the face recognition community. FRVT 2006 follows five previous face recognition technology evaluations - three FERET evaluations (1994, 1995 and 1996) and FRVT 2000 and 2002.

FRVT provides two datasets that are used for the purpose: high computational intensity test (HCInt) data set and Medium Computational Intensity test (MCInt) data set. HCInt has 121,589 operational well-posed frontal (within 10 degrees) images of 37,437 people, with at least three images of each person. The images are provided from the U.S. Department of State's Mexican non-immigrant visa archive. The images are of good quality and are gathered in a consistent manner, collected at U.S. consular offices using standard digital imaging apparatus whose specification remained fixed over the collection period.

The MCInt data set is a heterogeneous set composed of still images and video sequences of subjects in a variety of poses, activities and illumination conditions. The data are collected from several sources, captured indoors and outdoors, and include lose-range video clips and static images (with over hundred individuals), high quality still images, Exploration Video Sequences (where faces move through the nine facial poses used for the still images) and Facial Speech Videos (where two video clips were taken of individuals speaking, first in a neutral way, then in an animated way).

Face evaluation

For an evaluation to be accepted by the biometric community, the performance results have to published along with the evaluation protocol. An *evaluation protocol* describes how the experiments are run and how the data are collected. It should be written in sufficient detail so that users, developers, and vendors can repeat the evaluation.

The main attributes of the evaluation protocol are described below.

Image domain and face processing tasks

There are two image domains where Face Recognition Biometric Systems (FRBS) are applied:

1. *Face recognition in documents* (FRiD), in particular, face recognition from Machine Readable Travel Documents (MRTD), and
2. *Face recognition in video* (FRiV), also refered to as *Face in Crowd* problem, an example of which is face recognition from surveillance video and TV recordings.

These two image domains are very different [21]. The systems that perform well in one domain may not perform well in the other [18].

FRiD deals with facial data that are of high spacial resolution, but that are very limited or absent in **temporal domain**. — FRiD face images would normally have *intra-ocular distance* (IOD) of at least 60 pixels, which is the distance used in the **canonical face model** established by International Civil Aviation Organization (ICAO) for MRTD. There will however be not more than one or very few images available of the same person captured over a period of time.

In contrast, FRiV deals with facial images that are available in abundance in temporal domain but which are of much lower spatial resolution. The IOD of facial images in video is often lower than 60 pixels, due to the fact that face normally occupies less than one eighth of a video image, which itself is relatively small (352x240 for analog video or 720x480 for digital video) compared to a scanned document image. In fact, IOD of faces detected in video is often just slightly higher than or equal to 10 pixels, which is the minimal IOD that permits automatic detection of faces in images [25].

While for FRiD the facial images are often extracted beforehand and face recognition problem is considered in isolation from other **face processing** problems, FRiV requires that a system be capable of performing several other facial processing tasks prior to face recognition, such as face detection, face tracking, eye localization, best facial image selection or reconstruction, which may also be coupled with facial image accumulation and video snapshot resolution enhancement [20]. Evaluation of FRBS for FRiD is normally performed by testing a system on static facial images datasets described above. To evaluate FRBS for FRiV however, it is much more common to see the testings performed as a pilot project on a real-life video monitoring surveillance task [14], although some effort to evaluate their performance using prerecorded datasets and motion pictures has been also suggested and performed [22].

Use of colour

Colour information does not affect face recognition performance [27], which is why many countries still allow black-n-white face pictures in passport documents. Colour however plays an important role in face detection and tracking as well as in eye localization. Therefore, for testing recognition from video, colour video streams should be used.

Scenario taxonomy

The following scenario taxonomy is established to categorize the performance of biometric systems [26]: cooperative vs. non-cooperative, overt vs. covert, habituated vs. non-habituated, attended vs. non-attended, public vs. private, standard vs. non-standard. When performing evaluation of FRBS, these categories have to be indicated.

Dataset type and recognition task

Two types of datasets exist for recognition problems:

1. closed dataset, where each query face is present in the database, as in a watch list in the case of negative enrollment, or as in a list of computer users or ATM clients, in the case of positive enrollment,
2. open dataset, where query faces may not be (or very likely are not) in the database, as in the case of surveillance video monitoring.

FRBS can be used for one of three face recognition tasks:

1. *face verification*, also referred to as *face authentification* or *1 to 1 recognition*, or *positive verification*, as when verifying ATM clients,
2. *face identification*, also referred to as or *1 to N recognition*, or *negative identification* as when detecting suspects from a watch list), where a query face is compared against all faces in a database and the best match (or the best k matches) are selected to identify a person.
3. *face classification*, also referred to as *face categorization*, where a person is recognized as belonging to one of the limited number of classes, such as describing the person's gender (male, female), race (caucasian, asian etc), and various medical or genetic conditions (Down's Syndrome etc).

While the result of the verification and identification task are used as hard biometrics, the results from classification can be used as **soft biometrics**, similar to person's height or weight.

Performance measures

The performance is evaluated against two main errors a system can exhibit:

1. *false accept* (FA) also known as *false match* (FM), *false positive* (FP) or *Type I error*; and
2. *false reject* (FR) also known as *false non-match* (FNM) or *false negative* (FN) or *Type 2 error*.

By applying a FRBS on a significantly large data set of facial images, the total number of FA and FR are measured and used to compute one or several of the following cumulative measurements and figures of merit (FOM).
For verification systems:

1. FA rate (FAR) with fixed FR rate.
2. FR rate (FRR), or *true acceptance* rate (TAR = 1 - FRR), also known as true positive or *hit rate*, at fixed FA rate.
3. *Detection Error Trade-off* (DET) curve, which is the graph of FAR vs FRR, which is obtained by varying the system parameters such as **match threshold**.
4. *Receiver Operator Characteristic* (ROC) curve, which is similar to DET curve, but plots TAR against FAR.
5. *Equal error rate* (EER), which the FAR measured when it equals FRR.

For identification systems:

1. *identification rate*, or *rank-1 identification*, which is number of times when the correct identity is chosen as the most likely candidate.
2. *rank-k identification rate* (Rk), which is number of times the the correct identity is in the top k most likely candidates
3. *Cumulative Match Characteristic (CMC)*, which plots the rank-k identification rate against k.

The rates are counted as percentages to the number of faces in a databases. DET and ROC curves are often plotted using logarithmic axes to better differentiate the systems that show similar performance.

Similarity metrics, normalization, and data fusion

Different types of metrics can be used to compare feature vectors of different faces to one another. The recognition results can also be normalized. Proper covariance-weighted metrics and normalization should be used when comparing the performance results obtained on different datasets.

When temporal data are available, as when recognizing a person from a video sequence, the recognition results are often integrated over time in a procedure known as evidence accumulation or data fusion. The details of this should be given in the evaluation protocol.

Example protocols

Feret protocol [4] is an example of the *close set face identification*, where a full distance matrix that measures the similarity between each query image and each database image is computed. FRVT2002 [15] addresses both *open set verification problem* and *close-set identification problem* and uses CMC and ROC to compare the results. BANCA protocol [6], which is designed for multi-modal databases, is an example of the *open set verification protocol*. XM2VTS Lausanne protocol [10] is an example of a *close set verification*, where anyone not in the database is considered an imposter.

Evaluation Results

Face Databases have been used over the years to compare and improve the existing face recognition techniques. Some of the obtained evaluation results are shown in Figure 2. Figure 2.a shows face identification results from [16] for popular appearance-based face-recognition techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA), obtained on FERET database using CMC curves.

Figures 2.b-e show performance evaluation of commercial FRBSs that participated in the FRVT2002 and FRVT2006 tests taken from from [15] and [17].

Future work

Considerable advances have been made recently in the area of automated face recognition. FRBSs are now able to *recognize faces in documents* with the performance that matches or exceeds the human recognition performance. In large part, this has become possible due to the help of many researchers that have collected and maintained face databases. At the same time, despite the intensive use of these databases, no FRBS has been developed so far that can *recognize faces from video* with the performance close to that of humans.

Automated recognition of faces from video is considerably worse than face recognition from documents, whereas for humans it is known to be the opposite. This status-quo situation serves as an indication that new evaluation datasets and benchmarks are needed for testing *video-based face recognition systems*. With growing amount of video data easily accessible for public (including news casts, televised shows, motion pictures, etc.), it is foreseen that instead of using video-based data-bases, which are very costly and time consuming to create, the research community will soon adopt face evaluation benchmarks and protocols based on public domain video recordings [22].

The importance of improving the performance of video-based face recognition should not be underestimated, taking into account that of all *hard biometric* modalities, video-based face recognition is the most *collectable* and *acceptable* [24].

Related Entries

Face recognition, face detection, identification, verification, authentication, figures of merit.

References

1. In *Face Recognition website, http://www.face-rec.org*.
2. In *Face Recognition Vendor Test website, http://www.frvt.org*.
3. In *AT&T "The Database of Faces" (formerly "The ORL Database of Faces", http://www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html*.

4. In *P. J. Phillips, H. Moon, S. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):10901104, 2000. http://www.nist.gov/humanid/feret/.*
5. In *A. Georghiades, D. Kriegman, and P. Belhumeur. From few to many: generative models for recognition under variable pose and illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6):643660, 2001.*
6. In *E. Bailly-Bailliere, S. Bengio, F. Bimbot,M. Hamouz, J. Kittler, J.Mariethoz, J.Matas, K.Messer, V. Popovici, F. Poree, B. Ruiz, and J.-P. Thiran, The BANCA database and evaluation protocol. In Audio- and Video-Based Biometric Person Authentication (AVBPA), pages 625-638, 2003.*
7. In *P. J. Phillips. Human identification technical challenges. In IEEE International Conference on Image Processing, volume 1, pages 2225, 2002.*
8. In *A. OToole, J. Harms, S. Snow, D. R. Hurst, M. Pappas, and H. Abdi. A video database of moving faces and people. submitted, 2003.*
9. In *A. R.Martinez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center(CVC) Technical Report, Barcelona, 1998.*
10. In *K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: the extended M2VTS database. In Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.*
11. In *T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12):16151618, 2003. http://www.ri.cmu.edu/projects/project 418.html.*
12. In *B.-W. Hwang, H. Byun,M.-C. Roh, and S.-W. Lee. Performance evaluation of face recognition algorithms on the asian face database, KFDB. In Audio- and Video-Based Biometric Person Authentication (AVBPA), pages 557565, 2003.*
13. In *W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao. CAS-PEAL large-scale Chinese face database and evaluation protocols. Technical Report JDL-TR-04-FR-001, Joint Research and Development Laboratory, 2004. http://www.jdl.ac.cn/peal.*
14. In *R. Willing, Airport anti-terror systems flub tests face-recognition technology fails to flag suspects, in USA TODAY, September 4, 2003. Available at http://www.usatoday.com/usatonline/20030902/5460651s.htm.*
15. In *P. J. Phillips, P. Grother, J. M. Ross, D. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002: evaluation report, March 2003.*
16. In *K. Delac, M. Grgic, S. Grgic, Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set, International Journal of Imaging Systems and Technology, Vol. 15, Issue 5, pp. 252-260, 2006.*
17. In *Overview of the Face Recognition Grand Challenge - IEEE Conference on Computer Vision and Pattern Recognition, June 2005. Online at http://www.frvt.org/FRGC.*
18. D. Gorodnichy. Recognizing faces in video requires approaches different from those developed for face recognition in photographs. In *Proceedings of NATO IST - 044 Workshop on "Enhancing Information Systems Security through Biometrics". Ottawa, Ontario, Canada. October 18-20 (online at http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-47149.pdf*, 2004.
19. D. O. Gorodnichy. Facial recognition in video. In *Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA'03), LNCS 2688, pp. 505-514, Guildford, UK, online at http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-47150.pdf*, 2003.
20. D. O. Gorodnichy. Introduction to the First IEEE Workshop on Face Processing in Video. In *First IEEE CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, USA, online at http://www.visioninterface.net/fpiv04/preface.html*, 2004.
21. D. O. Gorodnichy. Video-based framework for recognizing people in video. In *Second Workshop on Face Processing in Video (FPiV'05). Proceedings of Second Canadian Conference on Computer and Robot Vision (CRV'05), pp. 330-338, Victoria, BC, Canada, ISBN 0-7695-2319-6, online at http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48216.pdf*, 2005.
22. D. O. Gorodnichy. Seeing faces in video by computers (Editorial). *Image and Video Computing, Special Issue on Face Processing in Video Sequences. (online at http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48295.pdf*), 24(6):1–6, 2006.
23. R. Gross. Face databases. In A. S.Li, editor, *Handbook of Face Recognition*. Springer, New York, February 2005.
24. A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(1):4–20, January 2004.
25. G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for realtime face detection and classification. In *Intern. Conf. on Automatic Face and Gesture Recognition, USA, pp 10-15*, 2002.
26. J. L. Wayman, A. K. Jain, D. Maltoni, and D. Maio, editors. *Biometric Systems: Technology, Design and Performance Evaluation*. Springer, New York, 2005.
27. A. Yip and P. Sinha. Role of color in face recognition. *MIT tech report (ai.mit.com) AIM-2001-035 CBCL-212*, 2001.

Definitional Entries

Face processing

Face processing is a term used to describe image processing tasks related to extraction and manipulation of information about human faces, such as face segmentation, face detection, face tracking, face modeling, face accumulation or fusing, face classification, facial expression recognition, face memorization and face identification. The term was originally introduced for the first IEEE workshop on Face Processing in Video held in 2004 [20], and is now applied for face processing in any sensory data.

Feature vector

Feature vector is a multi-dimensional vector that is obtained from a face by using feature extraction and image processing techniques to be used and that is used to memorize and recognize the face.

Large Scale Evaluation

Large Scale Evaluation is the evaluation that involves testing on significant amounts of data. It normally provides results using the statistical measurements such as average FAR, FRR and/or ROC and CMC curves.

Canonical face model

Canonical face model is the model that is used to store face images in databases. Once a facial image is acquired, it is resized and transformed to match the size and the orientation of the canonical face model, in which it is then stored in a database and used for face recognition tasks. For face recognition in documents, the canonical face model proposed by the International Civil Aviation Organization (ICAO) for the Machine Readable Travel Documents is used by many passport and immigration offices in many countries. This model stores faces using 60 pixels between the eyes, which ensures that feature-based face recognition techniques can be applied on these images. It has been argued however that this canonical face model may not be suitable for face recognition in video, due to the fact that face resolution in video is normally lower than 60 pixels between the eyes [18].

Temporal domain

When a face is captured over a period of time, as in video recording, it is often said that a facial image is available in temporal domain or that it has temporal resolution. In contrast, when only a single image of a face is available, as in a passport photograph, it is said that facial image is not available in temporal domain. In sensing data, a natural tradeoff is observed: either sensory data are of high spatial resolution or temporal resolution, but not of both at the same time. For example, an image of a face in a printable document if of high resolution, whereas faces observed live on TV are normally of very small resolution. As demonstrated by biological vision systems, recognizing an object that is observed in temporal domain (e.g. recognizing a face on TV) can be done just as efficiently or even more efficiently than recognizing the same object from a single high-resolution sample. For automated recognition systems however this is not the case yet.

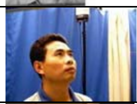
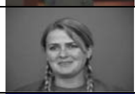
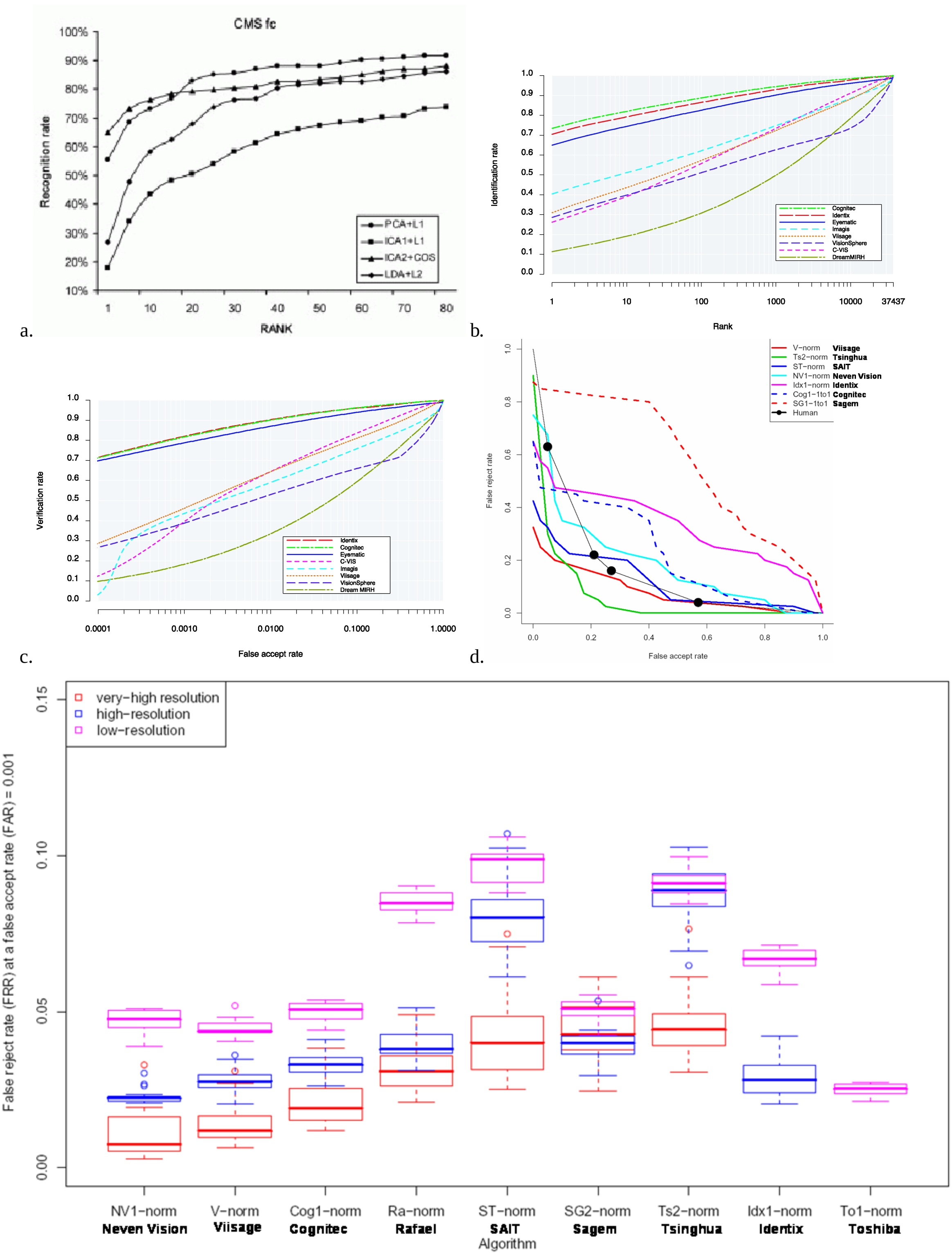
Database (year created) / Users	#individuals / # images	IOD / image width	Orientation	Expression	Lighting / quality	Occlusion	situations	Representative Facial image
AT &T Olivetti (1992-1994)	40 / 400	~60 / 92	yes	yes	yes	yes		
FERET (1993-1996)	1999 / 14,126	~ 80 / 256	9-20	2	2		2	
Yale (B)	15 / 165 10 / 5760	~80/ 640	9		64			
PIE 2000	68 / 41,368	~75/ 640	13	3	43			
AR >200	116 / 3288	~90 / 768	1	4	4	2 (eyeglasses, scarfs)	2	
Banca 2002-2003	208 / 208*12	~45 / 720	1	yes	3		12	
nist	573 / 3248	~80	2: front, profile					
Cas-peal 2003	1040	~45 / 360	21	15	6		1-5	
Notre-Dame Human ID	350 / 15.500	~80/ 1600	1	2	3		10	
U of Texas 2002	284	~80/ 720	video	video				
Korean	1000/ 52000	~80/ 640	7	5	16			
Equinox	91	~100 / 240	1	3	3 - IR images			
Cmu-hyperspectral	54	80/ 640	1		4 - IR images		5	
XM2VTS	293 /	~100 / 720	Full rotation	speaking		eyeglasses	4	
FRVT HCInt 1999-2002	37,437 / 121,589	>100	1				3	
FRVT MCInt 1999-2002	>100 63	>80, <80	several	Still and video	several		yes	

Fig. 1. Face databases categorized by the factors affecting the performance of face recognition systems such as: number of probes, face image resolution, head orientation, face expression, changed in lighting, image quality degradation, occlusion, and aging(situations).



e.

Fig. 2. Examples of performance evaluation conducted on face databases: a) identification performance of several appearance-based recognition algorithms from [16]) measured using CMC curves on FERET database, b-e) verification and identification performance of commercial face recognition biometrics systems on FRVT datasets (from [15, 17]), using CMC curves (b), ROC curve (c), DET curve (d) and fixed-FAR FRR distrubutions (e).