Evaluating the Complexity of Databases for Person Identification and Verification*

G. Thimm, S. Ben-Yacoub, J. Luettin

IDIAP, CP 592, CH-1920 Martigny, Switzerland

Abstract. Databases play an important role for the development and evaluation of methods for person identification, verification, and other tasks. Despite this fact, there exists no measure that indicates whether a given database is sufficient to train and/or to test a given algorithm. This paper proposes a method to rank the complexity of databases, respectively to validate whether a database is appropriate for the simulation of a given application. The first nearest neighbor and the mean square distance are validated to be suitable as minimal performance measures with respect to the problems of person verification and person identification.
Keywords: Person identification, person verification, database evaluation

1 Introduction

Computer vision systems are more and more based on statistical or heuristic methods. It is therefore important to compare alternative algorithms in order to evaluate their individual performance. As this is impossible in an analytical way, comparison is usually done by means of experiments. This requires the use of an identical database and an identical test protocol. The database, especially the test set, should be as close as possible to the real world conditions. If the test set is to easy, the algorithm will be overestimated, if it is too difficult, it will be underestimated. What one normally wants is to estimate the performance for the actual condition.

In practice, real-world data are often unavailable for legal and/or practical reasons. On the other hand, artificial datasets are often insufficient due to limited size, artifacts, similar illumination, similar position of subjects relative to the camera, the same background, similar facial expressions, or alike. Different databases have been used in face recognition which does not allow an objective comparison of results.

We describe therefore a procedure to rank databases. In the following, the term *complex* is used to compare two databases, although we are aware, that a proper, non-subjective definition in the mathematical sense can not exist. For example, the complexity of two databases is compared by means of the error rate yielded by a nearest neighbor algorithm.

* This work has been performed with financial support from the Swiss National Science Foundation under Contract No. 21 49 725 96 and the Swiss Office for Science and Education in the framework of the European ACTS-M2VTS project.

2 Ranking databases

Before a database for training and testing is created, the recording and definition of the test protocol has to be well designed. Parameters of a dataset like noise, number of classes, items per class, the amount of artifacts, control of illumination, head position, and so on, are important in this context and determine in some way the complexity of a database. It is surprising that only two recently registered databases define a test protocol: the Extended M2VTS database [13] and the FERET[1] [18] database.

Another, often neglected but still important, question concerns the reliability of obtained results. This reliability depends on the statistical significance of the test (see [21]) and how similar the evaluation database is compared to real world data. It is therefore desirable to rank datasets according to their complexity for a given task or problem $\mathcal{P}$.

Let $\mathcal{P}$ be problem defined for some class of objects. A computer can not directly act on physical objects, but on digital or analog signals. Such signals are obtained from applying transformations $\mathcal{T}_i \subset \{t_1, t_2, \ldots, t_p\}$ to the real objects, giving databases $\mathcal{D}_i$ (such transformations include for example the projection of the object to an image, as well as filtering in the computer). The transformations in $\mathcal{T}_i$ used for the production of $\mathcal{D}_i$ are directly related to the complexity of problem $\mathcal{P}$. Consequently, the goal is to rank the databases $\mathcal{D}_i$ according to a measure that reflects how well they incorporate the transformations that influence the complexity of a specific problem. However, datasets can not be ranked easily:

- The ranking depends on the problem $\mathcal{P}$.
- It is often unknown which transformations were applied to the objects and how well they reflect the transformation encountered in a real application.
- The assignment of a "degree of complexity" to specific transformations and to determine how complexities add up is difficult.
- Transformations are often continuous, implying different, continuous valued, degrees of complexity.

Given these unknown parameters, we propose to test a database my means of the performance of a gauge algorithm $\mathcal{A}$. The performance achieved by this algorithm on a particular database $\mathcal{D}_i$ is then used as the complexity measure of $\mathcal{D}_i$ with respect to $\mathcal{T}$ and $\mathcal{P}$.

3 Face Identification and Verification

In the context of this paper, the problems are $\mathcal{P}_1 =$ ***person identification*** and $\mathcal{P}_2 =$ ***person verification*** by means of faces (*i.e.* $\mathcal{T}$ operates on faces). The set of possible transformations that increase the complexity of problems $\mathcal{P}_{1,2}$ are a

[1] The FERET database is not publicly available. The authors have been unable to obtain it.

subset of $\mathcal{T} = \{$***rotation, illumination change, scaling, facial expression***$, \ldots\}$. The gauge algorithms are chosen to be the $\mathcal{A}_1 =$ ***nearest neighbor classifier***, respectively the $\mathcal{A}_2 =$ ***mean square distance*** (applied to zero-mean normalized image vectors). Obviously, both algorithms are not robust against illumination changes, translations, rotations, or scaling. Although other choices for $\mathcal{A}_{\{1,2\}}$ are possible, the chosen methods are most suitable for the following reasons:

- No free parameters have to be defined. Algorithms based on approaches like neural networks, genetic algorithms, and so on, require some parameters to be standardized (learning rate, network topology, crossover ratios,...).
- The first nearest neighbor algorithm and the mean square distance are well known and easily implemented and therefore cause only little work overhead.
- The complexity of the algorithms is reasonably low.

4 Experiments

The aim is (1) to evaluate the complexity of commonly used databases, (2) to compare identification and verification performance, (3) to compare the performance of the nearest neighbor algorithm, respectively the mean square distance, with published algorithms, and (4) to estimate for which evaluation tasks these datasets are appropriate. Scanning papers concerned with face recognition based on frontal views ([3, 4, 6–8] and others) reveals that many different datasets are used (table 1) - some not even publicly available - which prevents the comparison of published results. Some databases are available via [9]. Sometimes, mixtures of databases from different independent sources were used, in the aim to increase the significance of an evaluation [12, 23].

Name of the database	#
FERET [18]	11
Private or unspecified databases	11
ORL [19]	6
Mixtures of other databases	5
M2VTS [17]	4
Weizmann [14]	2
Yale Face Database [5]	1
Bern [1]	1
MIT [22]	1

Table 1. Databases used for person identification or verification and the number of times (column #) used in literature. The FERET database is often used in parts only.

Four datasets are used in this report (for sample images see figure 1). Only the Extended M2VTS database includes a well defined training and testing procedure:

1. The **Weizmann Institute of Science database** (subjects: 28, images per subject: 30) [14]. The images show the head, the neck, and some amount of background. The images are scaled to 18×26 pixels. The database was split twice into 50 pairs of training and test sets. The first 50 training sets included 8 images of each identity with the same, randomly chosen head positions and illumination. The second set of training sets contains also 8 images per identity, but not necessarily the same shots.

2. The **Bern database** (subjects: 30, images per subject: 10) [1]. The position and orientation of the faces is controlled, but the faces are neither centered nor scaled. As the rotation angles of the head positions are smaller as compared to the Extended M2VTS database, we decided to "crop" the images. In this operation, first top rows and columns with a high amount of background are removed (the hair remained mainly). Then, the lower part of the image is cut/extended, in order to obtain an image that has a height/width ratio of 2/3. Finally the images are scaled to 20×30 pixel.
   Two experiments were performed, each using 20 pairs of training and test sets. Each training set contains 4 randomly chosen images of each person. However, during the first experiments, the training sets contained always the same shots of each person.
3. The **ORL database** from the Olivetti Research Laboratory (subjects: 40, images per subject: 10) [19]. The faces in this database are already centered and show only the face. For the experiments, the images are scaled to 23×28 pixels, the database was split into 10 training and test sets. Each identity is represented 4 times in each training set.
4. The **Extended M2VTS database** (subjects: 295, images per subject: 8) [13]. The faces in this database are neither equally positioned, nor scaled. The faces are detected using the Eigenface algorithm [22], and the eyes are located using again the Eigenface approach. Then, the positions of the eyes are used to normalize the scale, to rotate the head into an upright position, and to define the region of interest. The region of interest is extracted, scaled, and stored as grey level image of the size 24×35 pixels. In a small percentage of the images the eyes were hand-labeled, as the head or eyes were not properly detected. The experiments were performed with six different pairs of training and test sets, each containing 4 images from two sessions.

Two tests were performed with these databases:

1. Person identification using a first nearest neighbor classifier with a mean square distance measure. The performance measure is the correct classification rate (see table 2).
2. Person verification using the mean square distance. The performance measure is the equal error rate (see table 3).

As the Weizmann and Bern databases are controlled (head position, illumination direction, and facial expression in the Weizmann database, the head position in the Bern database), two experiments were performed for each dataset. In the first experiment, with the results documented in the second column of table 2 and 3, the same shots of each identity were included in the training set. In the second experiment, all shots were selected randomly. It could, for example, occur that the training set for one identity includes only views of the left side of the face, whereas for another identity only frontal views are included.

The different performances show that small details in the configuration of an experiment may result in large changes of the error rate. Remarkable is also the high variance of the equal error rate when the training sets include always the

(a) Samples of the Weizmann database

(b) Samples of the Bern database

(c) Samples of the ORL database

(d) Samples of the Extended M2VTS database

Fig. 1. Samples from the four databases compared in this study.

Database	Average correct Identification		Number of Subjects
	similar shots	random shots	
Weizmann	85%	55%	28
Bern	85%	80%	30
ORL	—	92%	40
Ext. M2VTS	—	56%	295

Table 2. Percent of correct identification of the nearest neighbor classifier for face recognition (for the similar/random shots for all identities in the test set, if applicable).

same shot for each person. Equal error rates in the range of 5 to 13% for the Bern database and 9 to 33% for the Weizmann database have been observed. This demonstrates the importance of defining a common test protocol.

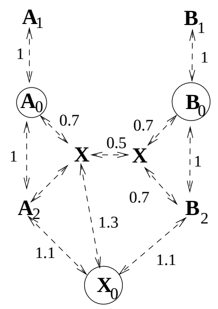
The high discrepancy of the identification rate and equal error rate, although not very intuitive, is explainable; see figure 2. Generally, this discrepancy is likely to occur when the inner- and inter-class distances are similar.

The results show, that, in terms of person identification, the ORL database has a lower complexity than the other three databases. The fact, that the number of persons contained in this database is higher than in the Weizmann and Bern database, supports the hypothesis, that the size of the database is not necessarily an indicator for the complexity of a database. Overall, the complexity does not increase with the size of the data set (as it would be expected).

For the problem of person verification, one would expect, that the average equal error rate is almost independent of the size of the data set (or slightly increasing with it). This is not true for the Weizmann and the Extended M2VTS

Database	Average Equal Error Rate		Number of Subjects
	similar shots	random shots	
Weizmann	16%	18%	28
Bern	10 %	11%	30
ORL	—	7%	40
Ext. M2VTS	—	11%(*)	295

Table 3. Equal error rate using the mean square distance for face verification (for the similar / random shots for all identities in the test set, if applicable). (*) Using the protocol described in [13] (without imposter accesses by clients), the equal error rate is 14.8%.



Consider the classes **A**, **B**, and **X** in a two dimensional space, with their elements distributed as shown. The dashed arrows indicate the euclidian distance between the elements, and circles indicate the test set. It can be easily seen that a nearest neighbor classifier has a 0% recognition rate. The equal error rate is 33% for a threshold of 1.2: in 6 tests, only $\mathbf{A}_0$ and $\mathbf{B}_0$ are accepted falsely for class **X**; in 3 tests, only $\mathbf{X}_0$ is falsely rejected. Note that the tree classes can easily be separated vertical lines.

Fig. 2. The performance of a identification and a verification system can be very different.

database: the latter includes a factor of 10 more identities, but the observed equal error rate is considerably lower.

It can be observed that the complexity of databases depends on the defined problem (*i.e.* for the Ext. M2VTS database). A database with a high complexity with respect to classification, is not necessarily complex with respect to verification, and vice versa. According to the experiments, the Extended M2VTS database is the most challenging of the four examined database for person identification and the Weizmann database for person verification.

5 Comparison with Other Publications

The nearest neighbor classifier obtains a considerable performance (table 4), when compared to other methods. Note that the ranking may change slightly due to different, in the respective papers often unspecified, test protocols. Unfortunately, the authors could not find any publications using one of these databases in the context of person verification.

In real applications, the nearest neighbor algorithm can in the presence of, for example, rotation and illumination changes **not** be expected to perform better

Weizmann database	
100%	Elastic matching [23]
85%	**Nearest neighbor**
84%	Eigenfaces [23]
~80%	Garbor-like filters [2]
41%	Auto-Association and Classification networks [23]

Bern database	
93%	Elastic matching [23]
87%	Eigenfaces [23]
85%	**Nearest neighbor**
43%	Auto-association and Classification networks [23]

ORL database	
96.2%	Convolutional neural networks [11]
95%	Pseudo-2D HMMs [20]
92%	**Nearest neighbor**
90%	Eigenfaces [22]
87%	HMMs [19]
84%	HMMs [16]
84%	Point matching and 3D modelization [10, 15]
80%	Eigenfaces [23]
80%	Elastic matching [23]
20%	Auto-association and Classification networks [23]

Table 4. Recognition rates reported in other publications.

than more sophisticated methods that take advantage from *a priori* knowledge. From the high performance of the nearest neighbor method, it can be concluded that the ORL database and probably the Bern database are insufficient for realistic tests of person identification applications.

6 Conclusion

It is argued and supported by experiments that it is necessary to rank databases used for the development and the comparison of classification and verification tasks. This helps to prevent a gross over- or underestimation of a system due to an inappropriate database. In consequence, a simple method is proposed that ranks databases according to their complexity prior to their usage. The argumentation is supported by experiments using four datasets and the nearest neighbor classifier for person identification, respectively a mean square distance measure for identity verification.

Among the four examined databases, the Extended M2VTS database is the most challenging database for person identification and the Weizmann database for person verification.

The first nearest neighbor method is shown to perform better than several other methods for person identification. Similarly, the mean square distance performs rather well for person verification on some databases. These outcomes and the simplicity of both approaches suggest to use these two methods as a minimal performance measure for other algorithms in their respective domains.

References

1. B. Ackermann: Bern data base (1995). Anonymous ftp: `iamftp.unibe.ch/pub/Images/FaceImages/`.

2. Y. Adini, Y. Moses, and S. Ullman: Face recognition: The problem of compensating for changes in illumination direction, IEEE Trans. on Pattern Analysis and Machine Intelligence 19 (July 1997) 721–732.
3. IEEE Int. Conf. on Automatic Face and Gesture Recognition, Killington, Vermont, IEEE (October 14-16, 1998).
4. IEEE Proc. of the Second Int. Conf. on Automatic Face and Gesture Recognition, Nara, Japan, IEEE (April 14-16, 1998).
5. P. N. Belhumeur and D. J. Kriegman: The Yale face database (1997). URL: `http://giskard.eng.yale.edu/yalefaces/yalefaces.html`.
6. J. Bigün, G. Chollet, and G. Borgefors, eds.: Audio- and Video-based Biometric Person Authentication (AVBPA'97), Lecture Notes in Computer Science 1206, Crans-Montana, Switzerland, Springer (March 1997).
7. H. Burkhardt and B. Neumann, eds.: Computer Vision - ECCV'98, II of Lecture Notes in Computer Science 1406, Freiburg, Germany, Springer (June 1998).
8. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR-96), San Francisco, California (June 18-20, 1996).
9. P. Kruizinga: The face recognition home page. URL: `http://www.cs.rug.nl/~peterkr/FACE/face.html`.
10. K.-M. Lam and H. Yan: An analytic-to-holistic approach for face recognition on a single frontal view, IEEE Trans. on Pattern Analysis and Machine Intelligence 20 (July 1998) 673–689.
11. S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back: Face recognition: a convolutional neural-network approach, IEEE Trans. on Neural Networks 8 (1997) 98–113.
12. S.Z. Li and J. Lu: Generalized capacity of face database for face recognition, in [4] 402–405.
13. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre: XM2VTSDB: The extended m2vts database, in Proc. Second Int. Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA'99) (1999). `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vts`
14. Y. Moses: Weizmann institute database (1997). Anonymous ftp: `ftp.eris.weizmann.ac.il/pub/FaceBase`.
15. A.R. Mirhosseini, H. Yan, K.-M. Lam, and T. Pham: Human face image recognition: An evidence aggregation approach, Computer Vision and Image Understanding 71 (August 1998) 213–230.
16. A. V. Nefian and M. H. Hayes III: Hidden markov models for face recognition, in ICASSP'98 5, IEEE (1998) 2721–2724.
17. S. Pigeon and L. Vandendorpe: The M2VTS multimodal face database, in [6].
18. P. Phillips, H. Wechsler, J. Huang, and P. Rauss: The FERET database and evaluation procedure for face recognition algorithms. To appear in: Image and Vision Computing Journal (1998).
19. F. Samaria and A. Harter: Parameterization of a stochastic model for human face identification, in Proc. of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota, FL (1994). URL: `http://www.cam-orl.co.uk/facedatabase.html`.
20. F. S. Samaria: Face Recognition using Hidden Markov Models. PhD thesis, Trinity College, University of Cambridge, Cambridge (1995).
21. W. Shen, M. Surette, and R. Khanna: Evaluation of automated biometrics-based identification and verification systems, Proc. of the IEEE 85 (September 1997) 1464.
22. M. Turk and A. Pentland: Eigenfaces for recognition, Journal of Cognitive Neuroscience 3:1 (1991) 71–96. ftp: `whitechapel.media.mit.edu/pub/images/`.
23. J. Zhang, Y. Yan, and M. Lades: Face recognition: Eigenface, elasic matching, and neural nets, Proc. of the IEEE: Automated Biometric Systems 85 (1997) 1423–1435.