IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Evaluating the Complexity of Databases for Person Identification and Verification

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

IDIAP–RR 98-10

revised in January 12, 1999

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 98-10

Evaluating the Complexity of Databases for Person Identification and Verification

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

revised in January 12, 1999

submitted for publication

Abstract. For the development and evaluation of methods for person identification, verification, and other tasks, databases play an important role. Despite this fact, there exists no measure whether a given database is sufficient to train and/or to test a given algorithm. This paper proposes a method to "grade" the complexity of a database, respectively to validate whether a database is appropriate for the simulation of a given application. Experiments support the argumentation that the complexity of a data set is not equivalent to its size. The "first nearest neighbor" method applied to image vectors is shown to perform reasonably well for person identification, respectively the mean square distance for person verification. This suggests to use them as a minimal performance measure for other algorithms.

Acknowledgements: 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.

1 Introduction

Overall in computer science, methods that are in some parts heuristic or guided by human intuition gain more and more importance. In order to show that these methods are valuable, they have to be evaluated on real world data. It is widely accepted - although unfortunately not rigorously done in the domain of person recognition or verification - that such evaluations have to be based on the same dataset and identical test protocols in order to obtain comparable results.

In practice, real-world data are often unavailable for legal reasons (*e.g.* a bank would not accept to take images of each person that uses an automatic teller machine) and practical reasons (the number of impostors is rather small as compared to regular accesses; restricted research budgets). Accesses to a system are therefore “simulated”, which results into a database that is often less difficult as compared to the real application (*i.e.* it has a lower complexity). This is not only due to the limited size, but also to artifacts as for example similar illumination, similar position of subjects relative to the camera, the same background, similar facial expressions, or alike.

It is therefore desirable to measure in some way the complexity of a given dataset in order to estimate whether it is appropriate for a system simulation of a given application.

2 Test Results are Biased by the Dataset

In this section, it will be argued that an inappropriate size or complexity of a dataset can lead to both, an over- or underestimation of the performance of a given algorithm.

Consider for example a neural network classifier (see [6] or [24] for introductions to neural networks). If the training set is rather small and corrupted by artifacts or noise, a neural network will not generalize well, and the optimal performance which would be obtained with a better training set might be underestimated. Note that a similar argumentation applies to other methods.

On the other hand, if the recording conditions of the database are too controlled (*i.e.* not realistic), a classifier might be unable to deal with “noisy” data as encountered in an application. The tests will therefore overestimate the system performance.

3 Protocols

As discussed above, the size and complexity of a database have a major influence on the results of an evaluation of a method that identifies or authenticates a person on the base of images. Furthermore, the protocol used for such an evaluation is important, *i.e.* which percentage is how selected for training and testing. Two recently registered databases take this into account and include instructions on how to split the data into training and test data: the Extended M2VTS database [14] and the FERET [18] database.

4 Grading a database

Databases are used to evaluate or to compare the performance of new techniques. An often neglected question that needs to be emphasized, is: how reliable are the obtained results? In other words, is the evaluation database complex (or difficult) enough in order to generate reliable results? It is therefore desirable to rank datasets according to the difficulty they represent for a given task or problem $\mathcal{P}$.

Let $\mathcal{P}$ be problem defined for some class of objects. However, a system does not directly act on the physical objects, but on representations in the form of images or numbers. Such representation are obtained from applying transformations $\mathcal{T}_i \subset \{t_1, t_2, \ldots, t_p\}$ to the real objects. The stored result are databases $\mathcal{D}_1, \ldots, \mathcal{D}_n$. The size and types of transformations in $\mathcal{T}_i$ used for the production of $\mathcal{D}_i$ define how difficult a problem $\mathcal{P}$ is.

The goal is to sort these databases according to a measure that reflects how well they incorporate the complexities that make a specific problem difficult. However, given datasets can not be ranked easily:

- The ranking depends on the problem $\mathcal{P}$.
- It is *a priori* not known which transformations were performed on the objects.
- The assignment of a "degree of difficulty" to a specific transformation is unsafe.
- Transformations are often continuous, implying different, continuous valued, degrees of difficulty.

The measure used to rank the databases has therefore a known parameter, the problem $\mathcal{P}$, and an unknown parameter, the set of transformations $\mathcal{T}$. We propose to test a dataset using an algorithm $\mathcal{A}$ which solves the problem $\mathcal{P}$ and is sensitive to the transformation set $\mathcal{T}$ (or a subset of it). Then, the negative performance achieved by this algorithm on a particular database $\mathcal{D}_i$ is the complexity measure of this database with respect to $\mathcal{T}$ and $\mathcal{P}$.

Example: Let $\mathcal{P}$ be the ***scale and rotation invariant object detection*** problem and the modality be the ***face*** (*i.e.* $\mathcal{T}$ operates on faces). The set of possible transformations that makes problem $\mathcal{P}$ difficult can be a subset of $\mathcal{T}$ = {***rotation, illumination change, scale change, facial expression,*** ...}. The gauge algorithm $\mathcal{A}$ can be the Eigenface approach [22], since it is sensitive to rotation and scaling. The performance of $\mathcal{A}$ on two databases permits to rank them according to their complexity. In other words, if the algorithm $\mathcal{A}$ performs well on a given database $\mathcal{D}_1$ and worse on $\mathcal{D}_2$, then $\mathcal{D}_1$ has a lower complexity (*e.g.* fewer scale, rotation, or illumination changes) than $\mathcal{D}_2$.

5 Face Identification and Verification

In the context of $\mathcal{P}_1$ = ***person identification*** or $\mathcal{P}_2$ = ***person verification***, the $\mathcal{A}_1$ = ***nearest neighbor classifier***, respectively the $\mathcal{A}_2$ = ***mean square distance***, applied to zero-mean normalized image vectors was chosen. 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 defined (learning rate, network topology, crossover ratios,...). As these parameters influence the performance of the whole system, they would have to be standardized.
- The first nearest neighbor algorithm and the mean square distance are easily implemented and therefore cause only little work overhead.
- The complexity of the algorithms is reasonable low.
- Both algorithms are well known.

The first nearest neighbor classifier applied to image vectors can be considered as non-robust, as illumination changes, translations, rotations, or scaling cause important differences for images of the same face.

6 Experiments

The aim of the experiments documented in this section is first, to support the hypothesis on the low dependence of dataset size and complexity, and secondly to demonstrate the approach in the domain of face recognition and verification.

Scanning papers concerned with face recognition based on frontal views (AFGR'98 [9], AFGR'96 [8], CVPR-96 [7], ECCV'98 [5], AVBPA'97 [4], and some other sources) reveals that many different and publicly not available databases were used, impairing the possibility to compare the results of the respective publications (see table 1). A selection of these and other databases is available via the **World Wide Web** [10]. In some publications, mixtures of databases from different independent sources were used, in the aim to increase the significance of an evaluation [23][13].

Name of the database	
Private or unspecified databases	11
Mixtures of other databases	5
FERET* [18]	11
M2VTS [19]	4
ORL [20]	6
Yale Face Database [3]	1
Weizmann [16]	2
Bern [1]	1
MIT [22]	1

*The FERET database was often used in parts only.

Table 1: databases used for person identification or verification and the frequency of their usage.

The following datasets are used in this report[1]:

1. The **Weizmann Institute of Science database** (subjects: 28, images per subject: 30) [16]. The images show the head, the neck, and a little bit of background (see figure 1). 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 with the same, randomly chosen head position 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]. Similarly to the Weizmann database, the position and orientation of the faces is controlled, but the faces are neither centered nor scaled (see figure 2). As the variation of the head positions are less important as compared to the Extended M2VTS database, we decided to "crop" the images. In this operation, first top rows and columns with an important amount of background are removed. Then, the lower part of the image is cut/extended, in order to obtain an image that has a height/width relation 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) [20]. The faces in this database are already centered and show only the face as shown in figure 3. 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) [14]. The faces in this database are neither equally positioned, nor scaled, as shown in figure 4. The faces are detected using an Eigenface algorithm, and the eyes are searched using again the Eigenface approach. Then, the positions of the eyes are used to normalize the scale, to rotate the head into an upright

[1] The authors could not yet obtain the FERET database.

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.

Figure 1: the Weizmann database

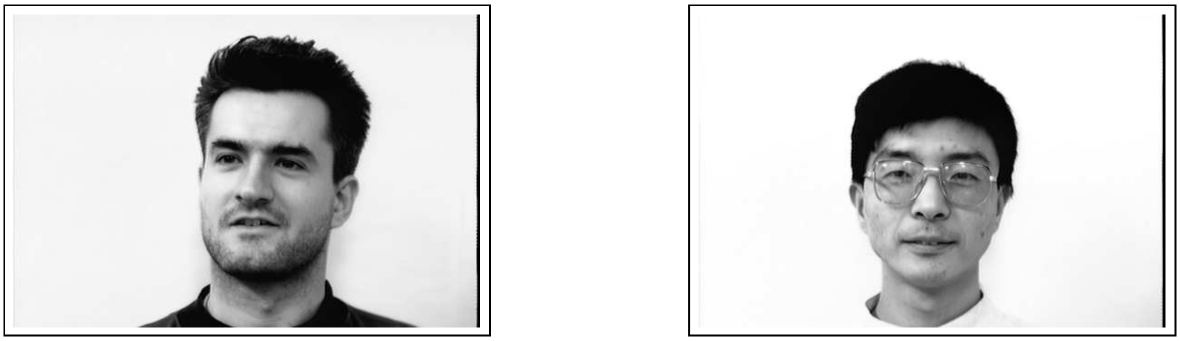

Figure 2: the Bern database

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 are known for the Weizmann database, the head position for the Bern database), two experiments were performed for each dataset. In the first experiment which corresponds to the first percentage in table 3 and 2, of all identities the same shots were included in into the training set. In the second experiment, the shots were selected randomly. It could, for example, occur that

Figure 3: the ORL database

Figure 4: the Extended M2VTS database

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 experiments using or using not the same shots for the Bern and Weizmann show that small details in the configuration of an experiment may result in important changes in the outcome. Remarkable is also the high variance of the equal error rate when the training sets include always the same shot for each person. Equal error rates in the range of $[5, 13]$ percent for the Bern database and $[9, 33]$ for the Weizmann database percent have been observed. Test protocols should therefore be described carefully in all details.

The high discrepancy of the identification rate and equal error rate, although not very intuitive, is explainable. Consider for example the classes $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{X}$ in a two dimensional space, with objects distributed as shown in figure 5. The dashed arrows indicate the distance between the elements, and circles the elements of the test set. It can be easily seen that the 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 $\mathbf{X}$; in 3 tests, only $\mathbf{X}_0$ is falsely rejected. Generally, this discrepancy is likely to occur when the inner-class distances are similar to the between-class distances.

The results show also that in the case of the ORL database, person identification and verification is much simpler than for 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

Database	Average correct Identification	Number of Subjects
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 same / different shots per identity, if applicable).

Database	Average Equal Error Rate	Number of Subjects
Weizmann	16% / 18%	28
Bern	10 % / 11%	30
ORL	7%	40
Ext. M2VTS	11%*	295

* Using the protocol described in [14] with non-client impostors, the equal error rate is 14.8%.

Table 3: equal error rate using the mean square distance for face verification (for the same / different shots per identity, if applicable).

size of the database is not necessarily an indicator for the quality of a database.

It can be seen, that the average equal error rate for the Weizmann database is higher than what is obtained for the Extended M2VTS database, and the equal error rate for the Extended M2VTS database and the Bern database are almost the same. This is in a stark contrast to the fact that the Extended M2VTS database includes a factor of 10 more identities.

An independence of the degrees of difficulties of the databases (*i.e.* for the Ext. M2VTS database) for person identification and verification can be observed. It can be concluded that a database that is difficult for classification, is not necessarily difficult for verification, and vice versa.

According to the working hypothesis, the Extended M2VTS database is the most challenging of the four examined database for person identification and the Weizmann database for person verification. Overall, the degree of difficulty does not increase with the size of the data set, neither is necessarily similar for closely related tasks like person identification and recognition.

7 Comparing with Other Publications

The nearest neighbor classifier obtains a considerable performance (table 4), where it is compared with 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 publication 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 than more sophisticated methods that take advantage from *a priori* knowledge. It can therefore be concluded that the ORL database and probably the Bern database are insufficient for a realistic application to person identification. On the other hand, table 4 shows that is it outperforms other, more complicated methods. It can therefore be concluded that, either the approaches are not appropriate for the problem, or the databases insufficient to evaluate them.

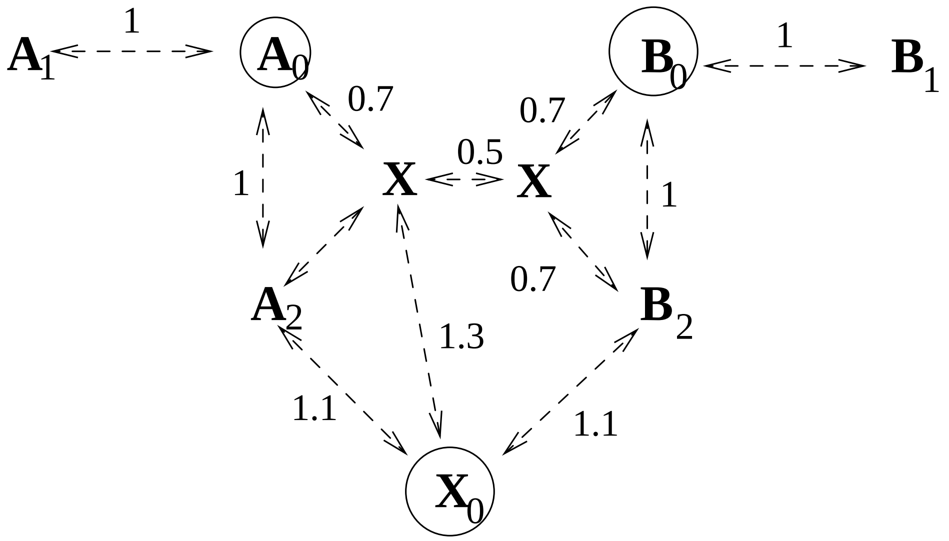

Figure 5: three classes with 0% recognition rate for the nearest neighbor algorithm and 33% equal error rate for the mean square distance.

8 Conclusion

It is argued that it is necessary to "grade" databases used for the development and the comparison of classification and verification tasks. Using an insufficiently complex database, where complexity is not equivalent to size, can result in an over- or underestimation of an algorithm. 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, 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] Bernard Ackermann, 1995, anonymous ftp: `iamftp.unibe.ch/pub/Images/FaceImages/`.

[2] Yael Adini, Yael Moses, and Shimon Ullman, *Face Recognition: The Problem of Compensating for Changes in Illumination Direction*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), no. 7, 721–732.

[3] Peter N. Belhumeur and David J. Kriegman, *The Yale Face Database*, 1997, URL: `http://giskard.eng.yale.edu/yalefaces/yalefaces.html`.

[4] 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.

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 [12]
95%	Pseudo-2D HMMs [21]
92%	**Nearest neighbor**
90%	Eigenfaces [22]
87%	HMMs [20]
84%	HMMs [17]
84%	Point matching and 3D modelization [11][15]
80%	Eigenfaces [23]
80%	Elastic matching [23]
20%	Auto-association and Classification networks [23]

Table 4: recognition rates reported in other publications.

[5] Hans Burkhardt and Bernd Neumann (eds.), *Computer Vision - ECCV'98*, Lecture Notes in Computer Science 1406, vol. II, Freiburg, Germany, Springer, June 1998.

[6] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Computation and Neural Systems Series; Santa Fe Institute Studies in the Sciences of Complexity; Lecture notes, vol. I, Addison-Wesley Publishing Company, The Advanced Book Program, Redwood City, California, 1991.

[7] IEEE, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-96)*, San Francisco, California, IEEE, June 18-20, 1996.

[8] IEEE, *International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, IEEE, October 14-16, 1998.

[9] IEEE, *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, IEEE, April 14-16, 1998.

[10] Peter Kruizinga, *The Face Recognition Home Page*, URL: `http://www.cs.rug.nl/~peterkr/FACE/face.html`.

[11] Kin-Man Lam and Hong Yan, *An Analytic-to-Holistic Approach for Face Recognition on a Single Frontal View*, IEEE on Pattern Analysis and Machine Intelligence **20** (1998), no. 7, 673–689.

[12] Steve Lawrence, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back, *Face Recognition: a Convolutional Neural-Network Approach*, IEEE Transactions on Neural Networks **8** (1997), no. 1, 98–113.

[13] Stan Z. Li and Juwei Lu, *Generalized Capacity of Face Database for Face Recognition*, In *Proceedings of the second International Conference on Automatic Face and Gesture Recognition* [9], pp. 402–405.

[14] J. Luettin and G. Maître, *Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB)*, IDIAP-COM 05, IDIAP, 1998, URL: `http://www.ee.surrey.ac.uk/Projects/M2VTS/`.

[15] Ali Reza Mirhosseini, Hong Yan, Kin-Man Lam, and Tuan Pham, *Human Face Image Recognition: An Evidence Aggregation Approach*, Computer Vision and Image Understanding **71** (1998), no. 2, 213–230.

[16] Yael Moses, *Weizmann Institute Database*, 1997, Anonymous ftp: `ftp.eris.weizmann.ac.il/pub/FaceBase`.

[17] Ara V. Nefian and Monson H. Hayes III, *Hidden Markov Models for Face Recognition*, ICASSP'98, vol. 5, IEEE, 1998, pp. 2721–2724.

[18] P.J. 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] S. Pigeon and L. Vandendorpe, *The M2VTS Multimodal Face Database*, Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science 1206, Springer Verlag, 1997.

[20] Ferdinando Samaria and Andy Harter, *Parameterization of a Stochastic Model for Human Face Identification*, Proceedings of 2nd IEEE Workshop on Applications of Computer Vision (Sarasota, FL), December 1994, URL: `http://www.cam-orl.co.uk/facedatabase.html`.

[21] Ferdinano Silvestro Samaria, *Face Recognition using Hidden Markov Models*, Ph.D. thesis, Trinity College, University of Cambridge, Cambridge, 1995.

[22] M. Turk and A. Pentland, *Eigenfaces for Recognition*, Journal of Cognitive Neuroscience **3** (1991), no. 1, 71–96, anonymous ftp: `whitechapel.media.mit.edu/pub/images/`.

[23] Jun Zhang, Yong Yan, and Martin Lades, *Face Recognition: Eigenface, Elasic Matching, and Neural Nets*, Proceedings of the IEEE: Automated Biometric Systems **85** (1997), no. 9, 1423–1435.

[24] Jacek M. Zurada, *Introduction to Artificial Neural Systems*, West Publishing Company, 1992.