Testing Face Recognition Systems

Graham Robertson* and Ian Craw
Department of Mathematical Sciences
University of Aberdeen, Aberdeen AB9 2UB

Abstract

Many papers can be cited which report results on face recognition techniques. Unfortunately the methods of testing and the data used for these tests vary considerably from paper to paper. This paper examines some the issues that should be considered when presenting face recognition results and when designing testing system. The paper promotes discussion of testing methods and suggests issues that researchers should consider and the types of information that should be included in their results.

1 Introduction

Face recognition by machine has long been recognised as a tricky task and over the years many suggestions as to how face recognition may be achieved have been made. In the early days Kaya and Kobayashi [6] used information theory to suggest that with 9 geometric measurements of the face it would be possible to recognise 92% of 5,000 faces. A recognition experiment, with manually located features, by Goldstein et al [4] reported a 50% recognition rate with around 250 different targets. Turk and Pentland [13] reported a 96% recognition rate for constrained images and 85% for images with unconstrained head orientations. Akamatsu et al [1] reported a recognition rate of 94% with about 400 trials and with a pool of twelve targets. Nakamura et al [8] claims 100% recognition of 10 cue images matched against a pool of 10 target images using an isodensity line technique. Sutherland et al [11] presented a technique using vector quantisation that achieved 89% recognition for a set of 600 pool images consisting of 30 people and 300 cue images containing images of the same 30 people. We have achieved 100% recognition with a pool of 45 images containing pictures of 3 different people and matching these images against another 60 test images of the same three people.

So what does all this mean? Who's results are the best? Is there any significance in these results? Sadly it is difficult to tell which if any of the results are good or useful. The three people in our pool of face images had significantly different hair colours which contributed greatly to the success of recognition process; yet previous results using the same methodology [3] were worse when the hair was included than when it was excluded. The number and types of constraints that researchers put on the data varies somewhat; e.g. to ensure that the subjects did not move their heads too much Kaya and Kobayashi [6] placed metal rods in

*GR is supported by SERC Research Grant Number GR/H75923

BMVC 1993 doi:10.5244/C.7.3

the subjects' ears. This paper is targeted at how to extract meaning from face recognition results, some of which are cited above. It examines the factors which affect results and identifies problems researchers face when trying to build a test set and report results. Some applications of face recognition are discussed because researchers often put their results in the context of their required application. In particular, we hope to encourage less emphasis on the 'headline' recognition result, and more discussion of the testing methodology and the chosen variability in the data.

We start with a brief description of face recognition whose main aim is to fix the language we adopt subsequently. Two categories of face recognition will be described, one based on geometric features and the other on texture features obtained directly from image intensities. Then follows a discussion of the difficulty in acquiring data and on the types of constraints that can be put on the acquired data. We will present results obtained under a variety of test conditions, showing that under certain circumstances the recognition is invariant to many lighting and head orientation conditions but not in others. In view of this testing, we consider the questions raised about some of our own earlier face recognition results, as well as those of others, in the final section.

2 Background

Terminology It is convenient to have a standard terminology with which to discuss recognition results. Typically, the task under discussion is to select one or more faces from a collection; we call this collection of face images the *pool*. The search is driven by the desire to match an example face, which we call the *cue*, while those members of the pool which are images of the same face as the cue are *targets*. Thus successful recognition selects one or more targets from the pool, while rejection will also report no match when there is no target in the pool, and so no *distractors* are recognised. Another collection of faces, which we call the *ensemble* is usually in evidence in existing systems, and is typically a way of providing background information about faces. Thus for recognition based on geometric techniques the general structure of faces and appropriate invariants are determined from the ensemble; in a vector quantisation approach the face components comprising the codebook are extracted from the ensemble; and in methods based on Principal Component analysis, the eigenface subspace, which again comprises the coding language, is derived from the ensemble. In some cases the pool and ensemble co-incide; but since this is not necessary, we try to keep the functional distinction even in these cases.

Neural net methods also map onto this language; the training set comprises the pool and the cues are taken subsequently from the test set, while the ensemble is used at an earlier stage, typically to select an appropriate architecture.

Recognition itself is often described in terms of rank ordering potential matches; with correct recognition occurring when the target heads the ranking. With many targets in the pool, a number of criteria can be used; best match, best average match amongst all sets of potential targets etc. We note that in some of our tests, not all methods give the same answer; it is clearly appropriate to lay down the criterion before starting the test! Finally we note the *censorship* problem; it is always possible to tune a pool by removing those faces which are easily confused with the target. It is essentially impossible to control for this; we report an example below where it would make a significant difference to the results, and in Fig 8, show why such misidentifications should occur according to current theories

of face recognition. We note also that in comparing methods, an unconscious bias by the originators may perhaps result in 'their' methods appearing to be best.

Types of recognition The fundamental task in face recognition is to develop a method of matching between target and cue, which depends on the identity of the faces themselves, and, within limits, is not influenced by imaging conditions such as pose and lighting or more subtle changes such as expression and age.

There are two main types of face recognition techniques. The first is recognition by face geometry where one tries to extract the distinctiveness of face shapes. The usual way of doing this is to select a set of landmarks on the pool of faces. The landmarks are often chosen through observations of an ensemble of faces and points on the face that are easily located such as the corners of the eyes and mouth are selected. The locations of the landmarks on cue faces are then compared to the locations in the pool faces and a matching criterion determined so that the target face will best match the cue. The advantage of this technique is the ease with which lighting invariance is achieved; clearly the landmarks on a bright image are in exactly the same locations as on a dark image. Pose and facial expression invariance cause more problems, because of movement of the relative locations of the landmarks. A solution may come from the use of a full 3D model of the human face, but at present we know of no system which implements such a method.

In fact even this assessment of geometric methods may be optimistic, since a useful system requires the automatic location of landmarks. Several systems have been proposed for doing this, including those of Kanade [5], Tock et al [12] and Robertson and Sharman [9]. They all start from the grey scale data, and use techniques such as edge-based analysis to 'understand' the face and hence locate the landmarks. And all the techniques proposed are prone to errors caused by changes in image acquisition conditions, including lighting variation!

Nevertheless there are two distinct problems here, and many researchers, including ourselves, perform recognition tests on images that have had the landmarks chosen manually in order to evaluate the potential of the method. Such results can then be compared with a truly automatic face recognition system. Indeed we would expect some advantages with the automatic system; we have experience with locations based on reliable feature detectors which suggest they can be more consistent than humans in locating landmarks, particularly when many data are to be collected for extensive testing. A second problem can arise when more than one operator is involved because of the difficulty in accurately specifying landmark locations.

A second recognition methodology, based on matching the grey levels on the cue and target faces, has been widely adopted recently. The intensities across a face image are related to the surface of the face in the image and the colour of the facial features. The method of comparing these images can simply be template matching, in which a cue image is correlated with all the faces in the pool; the image with the highest correlation is regarded as the best match. This method has produced high recognition rates in some circumstances (we achieved approx 86% in one test). Recognition is improved if all the faces used in the recognition process are the same size and are at a fixed location in the image. This can be achieved by placing crosshairs on the image acquisition device and lining up the eyes [7], nose and mouth or by performing an affine transform on the image to spatially normalise the face.

This simple technique can be improved upon by introducing principal component analysis [13] which extracts a face subspace. The subspace represent faces

well and extracts face characteristics that are good for recognition. The template matching technique can also be adapted by vector quantisation [11] which categorises the facial features into a set of features extracted from the ensemble. For a cue face the category for each feature is found by correlating the actual feature with sample features. By combining the category index for each feature a key is produced that can be compared with keys already known for the pool faces. The advantages with these image matching approaches to face recognition are that they are fairly simple to perform and facial orientation can be partially removed from the images by spatially and shape normalising the faces in the images. The main disadvantage is that the techniques are intolerant to changes in image acquisition conditions. More will be said about these problems later.

3 Acquisition of Face Data

Acquiring data is a hard and frustrating tasks for the face recognition researcher. It is even more frustrating given that the data can be seen walking around everywhere, and a face image can be grabbed very quickly. In fact we argue that such 'random' data sets can be confusing. Desirable properties include:

- the data should be as representative as possible of the population to be recognised;
- the data should be structured in such a way that the full range of acquisition conditions can be tested in a formal manner;
- enough data should be collected so that faces used for development differ from those used for testing;
- the data should be acquired over a significant period of time to allow the subjects faces to vary naturally;
- the face images should be of people that can be easily recalled for further tests as the research progresses; and
- the data should be made available freely so that others can compare results.

There is a a problem here as these requirements are inconsistent: a structured set of data is not representative of the real world population of faces; with data provided by others it is rarely possible to acquire new samples of the faces; and this problem also occurs when acquiring data over a lengthy period of time.

A further problem is that it is difficult to have a structured set of lighting conditions. It is easy to use photographic techniques to get good lighting on a subject but it is almost impossible to keep condition consistent throughout the test. Changes in conditions can be caused by small things such as a change in the size of the camera iris. This inability to control lighting has a significant effect on some recognition methods. If images are being acquired under incidental conditions good recognition may be achieved in the morning and bad recognition in the afternoon.

Further difficulties will appear as the methodolgy moves from the laboratory as the face databases in most citings have been specifically designed to exclude people with beards, moustaches, glasses and dark skin. This is obviously not representative of most populations to be recognised. If the system is to be used world-wide then ignoring faces from different cultures could also cause problems.

So how does one get a good set of data with which to work? Clearly different groups will settle for different compromises; we simply argue here for the importance of discussing the amount of data and the way they were collected when reporting recognition results. One interesting approach is by Akamatsu et al [1], who work with real 3D data from a head; they then manipulate this head using computer graphic techniques and hence obtain many samples in a controlled way. Even such an elaborate method has problems associated with constancy of expression etc.

4 Testing Recognition

In this section we describe a number of recognition results, with which we hope to put these criteria into perspective. In earlier work we have reported recognition over an age gap of 10 years, using the target shown in Fig 1 and a cue similar to those in Fig 2. In selecting targets and ensemble, we have conformed to some of the guide-lines above. Our ensemble and pool are disjoint, and have been chosen at random from the face database described by Shepherd [10], which contains 1000 faces taken in a standardised conditions, in the sense that a photographic studio was used and the subject was constrained to the point of using a neck clamp [10]. All this was done some ten years ago, so there is no question of interaction between methodology and image acquisition; the possibility of censorship remains, and we can simply state that to the best of our knowledge, none took place.

Despite the care taken, we now know of problems with the data; recently Cootes et al. were given landmark data for each of our set of 1000 faces, and performed a Principal Component Analysis of the resulting *shapes*, much as described in [2]. Had the attempt to standardise the data completely been successful, these components would have shown the variability associated with different face shapes; in fact two of the most prominent components (numbers 2 and 4) corresponding to the head nodding, and moving from side to side, indicating that even with the neck clamp in use the standardisation was not as complete as was hoped. For technical reasons, the landmarks were not those used for recognition, although we have we no reason to believe the results would have been any different.

Our data do fail other criteria described above. In particular we cannot extend the ensemble and pool in any way because without risking condition dependent recognition and so cannot report how our results change when imaging conditions within the pool and ensemble are allowed to vary. A substitute is to vary the conditions under which the cue is acquired, and we describe now such results. Our protocol was exactly as in [3], with an ensemble of 50, used to generate 21 Principal Components, and a pool of 100, including the target. As before we only report results with the hair removed; in this case the target remained the best match to the cue in each of the three conditions shown in Fig 2.

Of course others (eg Turk and Pentland [13] and Akamatsu et al [1]) have also reported tests that suggest that recognition via principal component analysis is tolerant to varying lighting conditions. At first these results seem promising but we now describe tests, consistent with all the above, in which the necessary invariance results fail to hold.

We took face images from five subjects, Nick, Graham, Peter, Robert and Grant. There are fifty samples of each of the first three faces and five samples each of Robert and Grant. The faces images were taken under several acquisition conditions including those labeled A and B below; we will call condition A 'plain' and condition B 'cluttered'. The image set contain faces faces with a range of

Figure 1: *Target*

Figure 2: *Three cues with varying conditions*

expressions. The conditions and expressions have no structure; we reproduce a sample of Grahams in Fig. 7 which give an indication of the variety of faces used.

We describe now an experiment which among others used the images shown in Figs. 3, 4, 5 and 6. The image backgrounds were not used in any tests, all the faces were spatially normalised, and the methodology was that in [3].

Figure 3: *Nick in condition A*

Figure 4: *Nick in condition B*

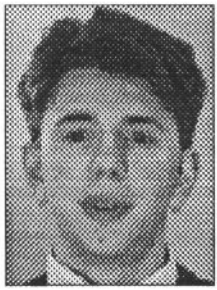

Figure 5: *Graham in condition A*

Figure 6: *Graham in condition B*

In one test, the plain Nicks (Fig. 3) and cluttered Grahams (Fig. 6) were selected and placed in the pool. Cueing another cluttered image of Graham matched the target to produce recognition; however, a plain Graham as cue was recognised as Nick, rather than Graham, so recognition appeared more influenced by condition than identity despite the fact that the backgrounds were ignored. In contrast, when the pool contained both Nick and Graham in the same (plain) condition, recognition occurred whether the plain or cluttered cue was used.

This experiment is similar to one performed by both Turk and Pentland [13] and Akamatsu et al [1]. Both reported that although the system was not trained with images under various lighting conditions recognition was good. A closer examination of these results may give and alternative explanation. In the second test described above, with a 'cluttered' Graham as cue, both subjects are poorly matched; however the match to the 'plain' Graham in the pool is better than that with the 'plain' Nick, and so recognition is successful. Yet in our first test above, the recognition process gave precedence first to the lighting conditions and only subsequently to the face characteristics. Other tests we have done show that varying facial expressions may cause similar problems to the those that occur with varying lighting conditions. Given a pool of faces with a neutral expression, and cue with a broad grin, we can get correct recognition of the target as described above. However, when a non target-face with a broad grin is introduced into the

pool, 'recognition' can follow the expression rather than the face.

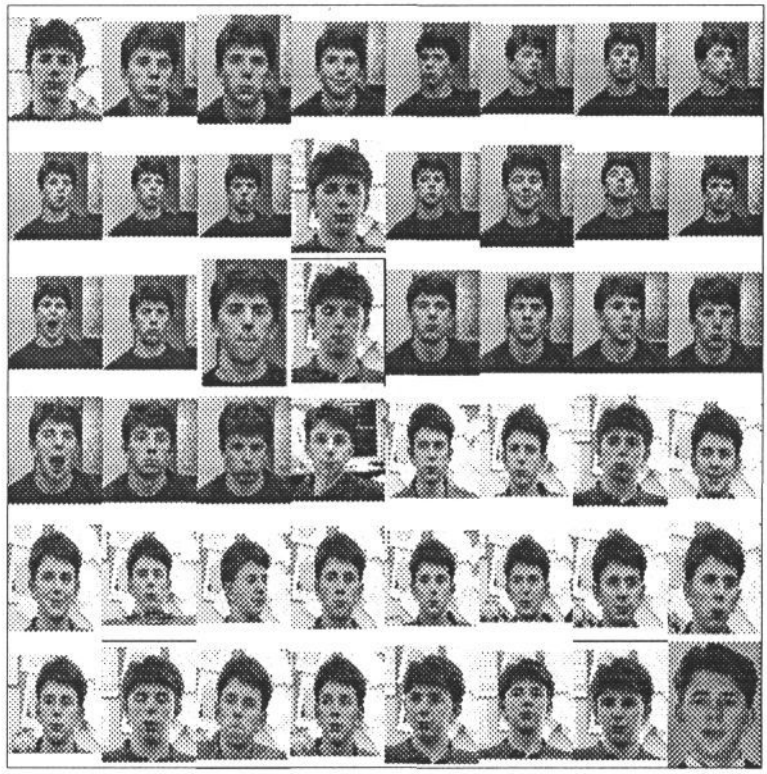

Figure 7: An example of the images used for the tests described in this paper. The faces are full face with unconstrained lighting and background conditions. The head is also allow to move around somewhat

5 Testing Rejection

Another important aspect of face recognition is the ability to recognise a single face, under relatively controlled conditions. The essence of such a system is the ability to reject unknown faces. Example applications include a security entry system with controlled lighting, in which a known position and orientation of the face, and a neutral expression, can be assumed; and a workstation security system that continually verifies that the user was the same person that logged on.

Rejecting unknown faces is currently an elusive goal, yet it is vital to such high security applications. All of the recognition results cited in the introduction were based on the closest match to a pool of faces. The results are therefore not even strictly recognition, but rather the selection, from a limited number of possibilities, of the most likely match, with no attempt at rejection if the match is bad.

Turk and Pentland addressed the problem of rejection in [13]. They first tightened the criterion for a matching face so much that only face images that were very similar to a pool faces were accepted and all others rejected. In this way, they were able to reject every distractor; however, despite the very standardised input

images the system also rejected valid personnel. Their application however did not involve 'one-shot' recognition, and they worked with few hundred pictures of a subject, and accepted recognition if at least one of these cues matched a member of the pool. In their tests, they reported success with this method, but more generally, the question still remains whether in such a circumstance the recognition process is adequately tolerant to inevitable changes in the imaging conditions.

Even such rejection results can however be hard to interpret. Using the principal component analysis technique described briefly above, with which a simple test achieved 100% recognition between three subjects Nick, Graham and Peter, we also tested the rejection performance. We tested the system with several images of an unknown face and found apparently reliable rejection; the system was able to detect that the face was not that of Nick, Peter or Graham. But in another test, a different face appeared to be consistently accepted as Nick; more certainly indeed than some of the cue images of Nick.

This phenomenon is at the heart of the rejection problem, and indeed can be predicted using a model rather like Valentine's model of face recognition [14], where the face is supposed to belong to some form of 'face space' regarded as a subset of $\mathbb{R}^n$. We illustrate this in the diagram below: the three similar discs represent different instances of the three faces clustering about some typical example in the face space $\mathbb{R}^2$; recognition is then achieved by assigning the appropriate 120° sector of the space to each of Nick, Graham and Peter. With this criterion, an unknown face, the dark disc, will be accepted as Nick. Indeed, if our discs are subject to error, and the elliptical regions are the 'correct' disc, the unknown face is always classified as Nick, while some examples of Nick are recognised as Peter; yet in this example, the ellipses themselves are disjoint and so can provide a basis for correct recognition.

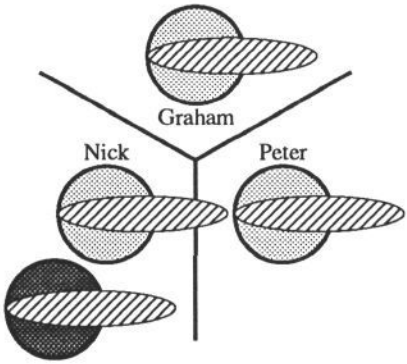

Figure 8: A simple 2D model of face space; the unknown (dark) face always lies in Nick's third of R^2, and so is recognised as Nick more often than Nick is.

6 Balanced Data

One approach when designing a recognition system is to train the system with many instances of each subject, varying expressions and lighting conditions in a consistent way. We shall refer to this as a 'balanced' pool: such a pool might contain a straight face, a smiling face and a face with a full grin for every subject.

The aim here is to achieve recognition within that range of conditions trained. Given the examples we have described, we argue that this must be done with care, since a good performance with a small set of subjects may not scale usefully. Two approaches are possible; either each combination of condition and face is treated as a separate instance, or all instances are combined leading to broad acceptance criteria (cf the disc in Fig. 8 is large). With broad acceptance criteria we expect rejection to be hard to establish, while when distinguishing face and condition, recognition itself may remain good with significant numbers of faces, providing the actual conditions used in the pool are reproduced quite closely, but the generalisation ability may not be useful, and in particular, a balanced pool may be essential to avoid recognition being influenced more by condition than identity.

The faces used by Turk and Pentland [13] form such a balanced set. This structuring of data is desirable in that it allows certain types of conditions to be tested to establish if those conditions are catered for well by the recognition technique. But to be sure that the results are valid the techniques should also be tested with a more random pool. This can be achieved by capturing pictures of faces that are moving around randomly or by increasing the size of the pool to such an extent that the structure cannot be maintained. The system can still be tested with the structured data as long as it is not related to the data in the pool.

7 Conclusions

This paper has discussed the testing of face recognition system. It has been shown that a system that works well with the restricted types of data or data that have been carefully structured may not necessarily perform as well under varied testing regimes. It is the intention of this paper to persuade researchers to examine more closely the recognition results they publish and to qualify them with more information about their testing process. More specifically to ask them to describe the constraints that there data were taken under. They should:

- comment on control of lighting conditions and facial expressions;
- indicate how closely the pool data are related to the testing data;
- show how many conditions are in the pool and test data;
- describe how structured the data are and how much variation is apparent in the subjects from whom the data are collected; and
- describe in detail any data used for testing the rejection of unknown faces.

The research field of face recognition has reached a point where apparently good recognition results can be achieved. The next stage is to investigate more systematically how the response varies when target, cues and distractors are obtained in a varying range of conditions. We have argued here that there can be difficulties in interpreting such invariance results simply, particularly when the conditions are carefully controlled and balanced; they must be put in context, and ideally they would be repeatable throughout the research community.

References

[1] Shigeru Akamatsu, Tsutomu Sasaki, Hideo Fukamachi, Nobuhiko Masui, and Yasuhito Suenaga. An accurate and robust face identification scheme. In *ICPR 92*, 1992.

[2] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Training models of shape from sets of examples. In David Hogg and Roger Boyle, editors, *British Machine Vision Conference 1992*, pages 9–18. Springer-Verlag, 1992.

[3] Ian Craw and Peter Cameron. Face recognition by computer. In David Hogg and Roger Boyle, editors, *British Machine Vision Conference 1992*, pages 498–507. Springer-Verlag, 1992.

[4] A.J. Goldstein, L.D. Harmon, and A.B. Lesk. Identification of human faces. *Proceedings of the IEE*, pages 748–760, May 1971.

[5] Takeo Kanade. *Computer Recognition of Human Faces*, volume 47 of *Interdisciplinary Systems Research*. Birkhäuser, Basel,Stuttgart, 1977.

[6] Y. Kaya and K.Kobayashi. A basic study on human face recogniton. In Satosi Watanabe, editor, *Frontiers of Pattern Recognition*, pages 265–289. Academic Press, London, 1972.

[7] M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for the characterisation of human faces. *IEEE: Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[8] O. Nakamura, S. Mathur, and T. Minami. Identification of human faces based on isodensity maps. *Pattern Recognition*, 24(3):263–272, 1991.

[9] G. Roberston and K.C. Sharman. Object location using proportions of the directions of intensity gradient - prodigy. In Canadian Image Processing and Pattern Recognition Society, editors, *Proceedings Vision Interface 92, Vancouver Canada*, pages 189–195, 1992.

[10] J. W. Shepherd. An interactive computer system for retrieving faces. In H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, editors, *Aspects of Face Processing*, chapter 10, pages 398–409. Martinus Nijhoff, Dordrecht, 1986. NATO ASI Series D: Behavioural and Social Sciences - No. 28.

[11] Ken Sutherland, D Rensham, and P B Denyer. A novel automatic face recogntion algorithm employing vector quantization. In *Colloquium: Machine Storage and Recognition of Faces. IEE Digest 017*, 1992.

[12] David Tock, Ian Craw, and Roly Lishman. A computer vision system for recognising and measuring facial features. Submitted to BMVC 91, May 1991.

[13] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[14] Tim Valentine. A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43A:161–204, 1991.