Ground truth and evaluation for latent fingerprint matching

Anna Mikaelyan and Josef Bigun
Halmstad University
SE-30118 Halmstad
{anna.mikaelyan,josef.bigun}@hh.se

Abstract

In forensic fingerprint studies annotated databases is important for evaluating the performance of matchers as well as for educating fingerprint experts. We have established ground truths of minutia level correspondences for the publicly available NIST SD27 data set, whose minutia have been extracted by forensic fingerprint experts. We performed verification tests with two publicly available minutia matchers, Bozorth3 and k-plet, yielding Equal Error Rates of 36% and 40% respectively, suggesting that they have similar (poor) ability to separate *a client from an impostor in latent versus tenprint queries. However, in an identification scenario, we found performance advantage of k-plet over Bozorth3, suggesting that the former can* rank *the similarities of fingerprints better. Regardless of the matcher, the general poor performance is a confirmation of previous findings related to latent vs tenprint matching. A finding influencing future practice is that the minutia level matching errors in terms of FA and FR may not be balanced (not equally good) even if FA and FR have been chosen to be so at finger level.*

1. Introduction

To evaluate methods for recognizing fingerprints, data sets as well as ground-truth are necessary. Example evaluations comprise feature extraction methods, features matching methods or imaging methods.

Fingerprint databases for evaluating methods are difficult to obtain for a variety of reasons. Firstly, there are legal restrictions, e.g. ID protection. Secondly they often have poor quality, e.g. for traces of individuals at crime scenes. As for the ground-truth, fingerprint experts, rather than image processing researchers, must annotate the valuable image features. In turn this demands considerable resources to construct such data sets.

Therefore, the fingerprint database provided by NIST (SD27, discussed below), which is annotated by fingerprint experts is an important resource for image analysis studies on fingerprints [1]. In this article, we reveal novel information on SD27 including establishment of minutia level correspondence of the ground truth, enlarging its use. Furthermore, we present performance of fingerprint matching on SD27 and discuss the forensic issues, by means of two publicly available minutiae matching techniques, Bozorth3 matcher [2] and k-plet matcher [3].

2. Database

The NIST SD27 data set has been proposed to develop novel methods, to assess existing systems, train human examiners on fingerprints, and promote standards, by NIST and FBI jointly (USA), [1], which also made it publicly available. It contains 258 pairs of fingerprints at 500 dpi resolution. Each pair consists of two images produced by the same finger, albeit at different times–a *tenprint* and a *latent*.

The tenprint is a good quality image, often imaged by rolling[1] a finger from nail-to-nail containing a rich set of minutia, and all cores and deltas (jointly referred to as singularity points, SP). The quality in this context refers to the ability of fingerprint examiners to extract identification information (minutia, SPs, orientation maps, etc). The high quality of tenprints is a result of them being recorded in a controlled manner. As a result, the variance in quality of tenprints is low.

By contrast, the discovery and imaging of latent fingerprints in a crime scene is a challenge in itself, demanding long training and experience. Despite advanced methods, the latents have significantly lower quality with higher variance (of quality) compared to tenprints. In SD27 the fingerprint experts use 3 categories to classify the quality of the latents–*good, bad, and ugly*. To be explicit, a good quality latent, Fig. 1 (top right), is still poor in comparison to a tenprint (top left), but nevertheless is more useful than an ugly fingerprint in identification, (bottom right).

[1] For many years this has been done by rolling the inked finger on cards.

Our first finding on SD27 concerns a double use of a tenprint. Although there are 258 different pairs, two pairs (G034 and U270) are special because they are from the same finger. All other pairs are from different fingers. The two tenprints in G034 and U270 are identical images but the latents are different images (of the same finger), one being Good the other being Ugly in quality. We have shown sub-parts of them in Fig. 1, top row and bottom right. In verification tests it is legitimate to use them as they are, i.e. different pairs, but care should be taken when counting ranking errors in identification tests.

For each fingerprint (both for a latent and a tenprint) there are two minutia sets, called the *Ideal* set and the *Matched* set. The minutia in the *Ideal* set of a latent were extracted by human experts without seeing the corresponding tenprint first, while those in the *Ideal* set of the tenprints were extracted automatically. The number of minutia in the *Ideal* set of a latent is much smaller than that of its client tenprint. As such this represents a condition close to the current practice of fingerprint examiners, justifying the notion of *Ideal*.

The minutia in the *Matched* set of a tenprint is a subset of the *Ideal* set. Similarly, the *Matched* set of a latent is a subset of the corresponding *Ideal*. The minutia of the *Matched* sets of a client pair (latent and its tenprint) correspond, which is also one of the most important assets of SD27. Hence, the number of minutia in the *Matched* set of a latent and that of the client tenprint agree. Surprisingly however, the correspondence is at the (*Matched*) set level. Although a human expert can obviously see the correspondence by viewing the two sets side by side, the pairwise correspondence of the minutia is not accessible to computers. In Fig. 2 (bottom left), we show the latent, and its client tenprint (top left), along with the associated *Matched* minutia sets overlayed and labeled with numbers corresponding to their indices extracted from the respective files of *Matched* sets.

It is clear that the indices provided by SD27 do not represent minutia labels and one of the *Matched* sets of a client pair must be permuted so that the minutia level ground truth of correspondence becomes accessible to computers. Fig. 2 (top-right) shows the relabeling of the native indices of SD27 for the *Matched* set of the latent (bottom-left) such that the novel indices represent the ground truth for the tenprint (top-left). How such ground truths are established for all 258 latent-tenprint pairs is discussed in Section 4.

3. Performance at finger-level

Identification is a semi-automatic process in forensics of fingerprints. Essentially, a human expert marks the minutia locations (possibly along with SPs) and their orientations in the latent and asks an Automatic Fingerprint Identification System (AFIS) to provide a ranked-list of tenprints that are ordered according to their resemblance to the latent. The query data consists in a sparse set of vectors, usually three dimensional integers–two for the x,y coordinates of the minutiae (or SP) location and one variable for the direction, e.g. see Fig. 1 bottom right. Thus, the matching of AFIS is based on minutia constellation. By contrast, the human expert uses her experience and her vision when comparing the tenprints (minutia vectors *and* images), pulled out by AFIS possibly among millions of tenprints, to the latent at hand (image and minutia). The expert either rejects the ranked-list all together or decides that one fingerprint in the list is the same as the latent and provides a value of evidence, a likelihood.

Although it is the human expert who actually carries the final responsibility of the decision on identification, the matching skills of an AFIS can influence the human performance negatively because the ranked-list may or may not contain the matching tenprint. It is worth to note that AFIS can be used not only for purposes of forensics, (latent-tenprint matching) but also in many other identification tasks, e.g. in visa applications where the query image is of high quality (tenprint-tenprint matching).

Figure 2 shows the FA and FR rates of match queries using two published minutia matching methods, Bozorth3 [2] and k-plet matcher [3], on SD27. Both methods output a similarity score if two (already extracted) minutia sets are presented as input. Since minutia sets are available for every tenprint and latent in the database, we could use the matchers to assess the similarity of an arbitrary latent against a tenprint. We used the *Ideal* set, in the above evaluation.

Given a latent, the similarity it has with its client tenprint should ideally be larger than if it is compared to an impostor tenprint. Both minutia matchers reported results that are consistent, amounting to a poor performance on the statistics of scores. The latter is summarized by score distributions, and more compactly by Equal Error Rate (EER), which was close to 40%. EER is a measure that represents how well a method separates the client matches from impostor matches in 1:1 matches, by finding the critical score of the system, and estimating the errors using it as a threshold in a decision rule. To be precise, the EER decision rule causes the impostor matches are falsely accepted (FA) at the same rate as the clients are rejected (FR), i.e. FAR=FRR=EER.

There is another way of summarizing the score statistics, often referred to as Cumulative Match Characteristic (CMC) curve for one to many matches (1:N), using ranks. It is adapted to identification engines, which suggest R candidate tenprints corresponding to a query (latent) searched against N tenprints. However, we studied matching methods which are verification engines i.e. they give a score on the similarity of two minutia sets. A verification engine can nevertheless be used to check a latent against a

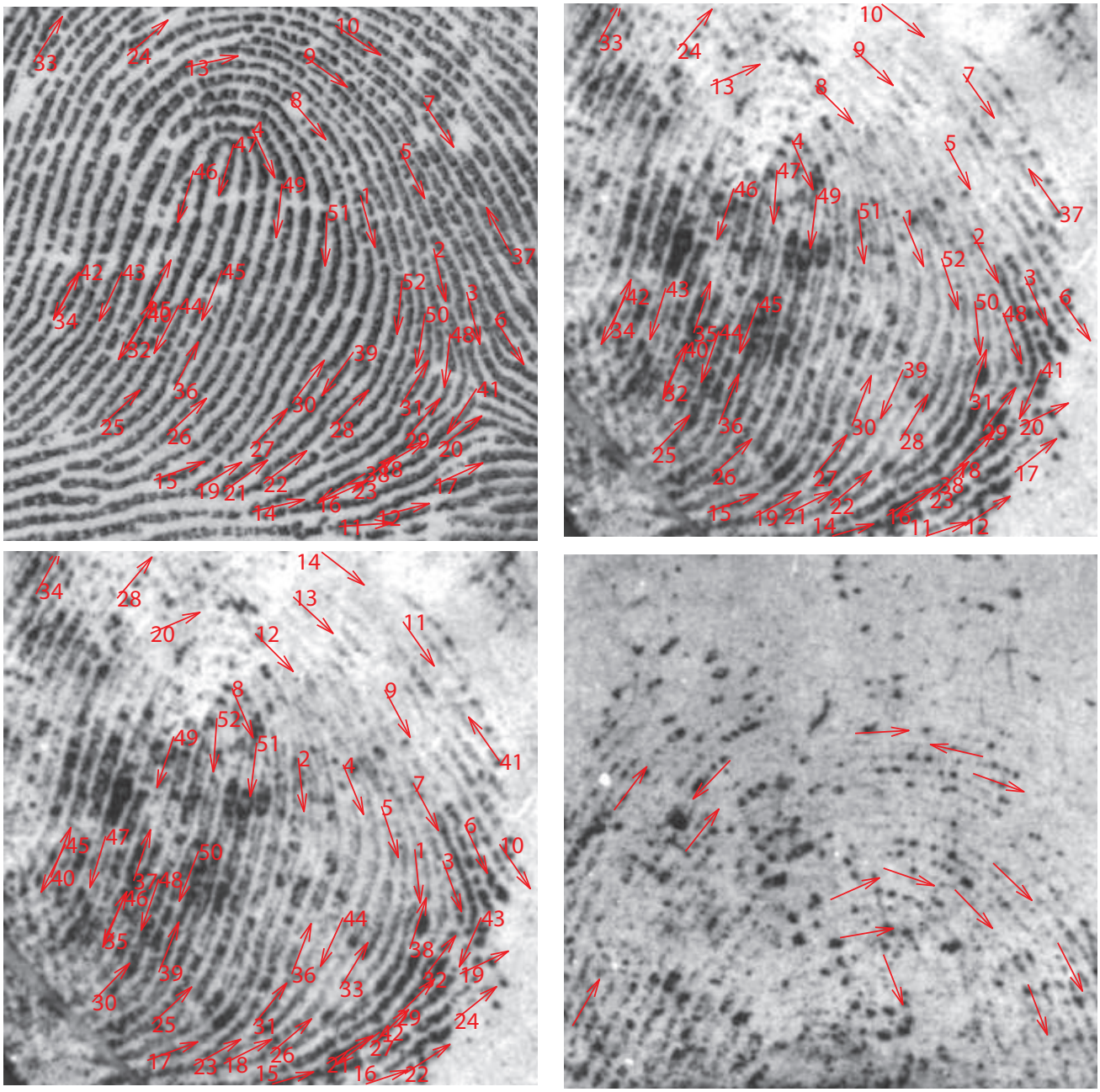

Figure 1. *(Top left)* A tenprint fingerprint with its minutia of *Matched* set overlayed. *(Top right)* A latent fingerprint that corresponds to the tenprint and that is quality-classified as **Good**. The minutia labels represent ground truth for minutia correspondence. *(Bottom left)* The same latent but the minutia labels, suggested by indices of SD27, do not represent the ground truth. *(Bottom right) A latent fingerprint that is classified as* **Ugly**

database consisting of N tenprints, one at a time, producing a score for each assessment. The scores can subsequently be ranked, and if the client tenprint is among the top R-ranked tenprints the query is judged as successful. Given a set of latents, and a large database of tenprints, containing the corresponding tenprints, one can change R to plot the frequency of successful searches, yielding a CMC curve. Evidently FA, FR and CMC curves depend on each other if the scores are produced by the same engine, [4]. CMC curves are useful to evaluate the ranking ability of a verification engine when used as an identification engine. Fig. 2 shows these curves for the two methods that we studied.

In a successfull query of a latent against a database that returns R tenprints with scores, the latent is matched against

$R-1$ (tenprints of) impostors and against one (tenprint of) the client. Considering that R will change (to produce the CMC curve), R will be limited by $R_{\max}$, i.e. $1 \leq R \leq R_{\max}$. Here, usiing SD27, we matched every latent against every tenprint, i.e. $R_{max} = 258$, to produce Fig. 2. However, the figure reports R up to 20 for readability.

Studying $R = 20$, one can see that the Bozorth3 method was successful in 66% of the queries, and k-plet method in 78% of the queries. Likewise, the rates of placing the client tenprint at the top rank, $R = 1$, and up to the second rank, $R = 2$, were 47 %, 50% and 55%, 60%, for the respective methods.

That the minutia constellation based latent identification performance is poor is also reported in [5] which suggests that in no more than 35% of the latent queries of SD27, their matching method was successful to place the client tenprints of the latents at the top-rank. The 12-20% lower performance compared to our reporting, can be explained by that they used a different matcher than ours and did not use the annotations provided by NIST DB for tenprints, but relied on those extracted by the (proprietary) software of Verifinger (company), i.e. the tenprint minutia data were different than ours. Another, difference is that they added a second unannotated DB, consisting of tenprints, to the tenprints set of DB27. Because they were unannotated it seems that they used a software to extract the minutia for all tenprints (including that of SD27). Finally, it is also not clear if they used the *Ideal* set of latent minutia as we did, since there is another minutia set available for each latent: the *Matched* set. We wish to stress that the relevant conclusion is that minutia constellation based fingerprint matching has ample room for performance improvement, e.g. by including additional features such as SPs, quality maps and orientation maps [5]. That using the latter leads to performance gains was previously suggested by another study as well [6] albeit in the context of tenprint-tenprint matches.

The difficulty of latent versus tenprint matches have also been confirmed by a NIST workshop on performance of commercial AFIS regarding latent queries against tenprints [7]. The SD27 latents were "pulled" out at 56% rate as the first rank, among 40 million tenprints. The poor AFIS performance is an additional justification for why the human fingerprint examiner's contributions in latent matching is indispensable. Though reduced, there is still a non-negligible risk that the automatic method will reject to include the client-tenprint into the ranked-list (e.g. 22-34% using SD27). This type of error, which is entirely to blame the machines for, is tolerated more than falsely accepting a tenprint as belonging to a client latent. The latter may find an innocent guilty whereas the former may contribute to a criminal escaping judgement. The lower the EER (or the top-rank rate (1:m)) the more efficient the output of the human expert will be.

4. Performance at minutia-level

Suppose that a machine matcher has verified a latent and a tenprint using the similarity between their minutia constellations. The matcher has then identified a subset of the latent minutia that it has matched with that of the tenprint at minutia level. Let each of these two subsets be the *Supposedly Matched* set of the respective fingerprint. How well the *Supposedly Matched* set agrees with the (true) *Matched* set is then of interest. To evaluate and improve a matcher, we thus need to know the ground truth of correspondence at minutia level, not only at finger level.

In the first phase of ground truth establishment, we used automatic matching. Initially, there were 5460 minutia in total in the *Matched* sets of the latents and as many in those of the tenprints, annotated by fingerprint examiners of FBI. To find the true correspondences, we added instructions to the source code of k-plet method, so that it would also produce its *Supposedly Matched* sets, since this was not its normal behavior. Subsequently, we used this to obtain all (258) *Supposedly Matched* sets. The automatic matching produced then 4672 correspondences rejecting the remainder, not found in tenprints although existing.

In the second phase of the ground truth establishment we compared and inspected the minutia visually, using the *Supposedly Matched* sets as starting point. For this we had to write displaying and editing software, overlaying the minutia on fingerprints with their new labels. The erroneous correspondences as well as missed correspondences were then possible to identify and correct by human intervention. We could note that a tiny fraction of the minutia (namely 11 of 5460) had to be deleted, both from latents and tenprints, because they contained obvious human errors of annotation[2]. These few errors were due either to impossible positioning of minutia within the respective constellations, that otherwise matched (5 cases), or the minutia directions were conflicting (6 cases) with 180 degrees[3].

After the second phase, the *Permuted Matched*[3] set was thus obtained for 258 pairs of SD27. This set contains 5449 visually verified, and ordered minutia such that the identities of the minutia, encoded in their storage order, correspond. Even when ignoring the 8 latent-fingerprint pairs containing erroneous minutia[2] entirely, the automatic matcher falsely rejected (by omission) 787 truly existing correspondences of 5224 total (15%). At the same time it suggested 170 minutia (3%) correspondences that were false. In total there were thus 954 minutia (18%) whose

[2]From fingerprints pairs labeled as G038, G044, G078, G080, U216 one minutiae was deleted from both latents and tenprints. From G082, G084, U242 two minutia were deleted from both latents and tenprints.

[3]To download our ground truth findings for SD27 see http://www.hh.se/staff/josef.

correspondences were not possible to establish on the basis of the minutia constellations automatically. In this part of our study, the matcher was given latent-tenprint pairs from the *Matched* sets, meaning that the problem was easier than the one experienced in operational conditions and hence the matcher should be more successful. This was also the case, because the EER obtained on the *Matched* set was 6 % (*Ideal* set 40 %).

Hence, the performance of a matcher degrades significantly if the number of minutia differ between the evaluated minutia sets, even if all latent minutia truly exist among those of the client tenprint. If one chooses a score threshold yielding a balanced verification error at finger level, e.g. 6% EER, this does not necessarily mean that the underlying minutia matches have balanced errors, here 3% FA with 15% FR. Accordingly, an AFIS may have low erroneous minutia associations between latents and tenprints, coming at the cost of a too frequent rejections of the true associations.

We were not able to obtain the minutia level performance of the Bozorth3 method because to deliver the *Supposedly Matched* set of an evaluation (a latent against a tenprint) is not in the normal behavior of the method. Neither was it practicable to modify the available source code, with the resources available to us.

5. Conclusion

Beside estimating scores for correspondences at finger level, fingerprint matchers can be used to establish minutia correspondences. To evaluate the performance of matchers in the latter task, annotated data-sets containing minutia level ground truths for correspondences are needed.

For forensic studies the *Permuted Matched* set of the SD27 is established. It represents the ground truths of correspondence at minutia level for minutia that exist both in latents and tenprints (the *Matched* set). The minutia in latents were extracted by forensic fingerprint experts.

The finger level matching performance degrades significantly if the number of minutia differ between the latent and the tenprint being evaluated. The minutia level matching errors in terms of FA and FR may not be balanced even if an operation point causing a balanced FA and FR at finger level is chosen. Consequently, the success of an AFIS in terms of low false minutia association rate between latents and tenprints may come at a too frequent rejection of true minutia association.

We studied the verification abilities of two publicly available minutia matchers, Bozorth3 and k-plet, yielding similar poor EERs, $\approx 36\%$ and 40%. By contrast, in an identification tasks the k-plet method faired better than Bozorth3, suggesting that the former can *rank* the similarities of fingerprints more reliably whereas the latter is slightly better in separating the impostor queries from client queries. Regardless of the matcher, the general poor performance is a confirmation of previous findings related to latent versus tenprint matching justifying further research.

References

[1] M. Garris and R. McCabe, "Nist special database 27: Fingerprint minutiae from latent and matching tenprint images," NIST, Gaithersburg, MD, USA, Tech. Rep., 2000. 1

[2] C. Watson, M. Garris, E. Tabassi, C. Wilson, R. McCabe, S. Janet, and K. Ko, "The nbis-ec software is subject to us export control laws." NIST, Gaithersburg, MD, USA, Tech. Rep., 2007. 1, 2

[3] S. Chikkerur, A. Cartwright, and V. Govindaraju, "K-plet and coupled bfs: a graph based fingerprint representation and matching algorithm," *Advances in Biometrics*, pp. 309–315, 2005. 1, 2

[4] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the roc curve and the cmc," *Automatic Identification Advanced Technologies, IEEE Workshop on*, vol. 0, pp. 15–20, 2005. 3

[5] A. Jain and J. Feng, "Latent fingerprint matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 88–100, 2011. 4

[6] H. Fronthaler, K. Kollreider, and J. Bigun, "Local features for enhancement and minutiae extraction in fingerprints," *Image Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 354–363, 2008. 4

[7] B. S. Swann, "Needs and applications of latents at fbi/cjis," in *Latent Testing Workshop 2006 Presentations*, V. Dyornychenko and M. D. Garris, Eds. NIST, april 2006. [Online]. Available: http://biometrics.nist.gov/cs_links/latent/workshop06/proc/P6_Swann_LatentOverview_2.pdf 4

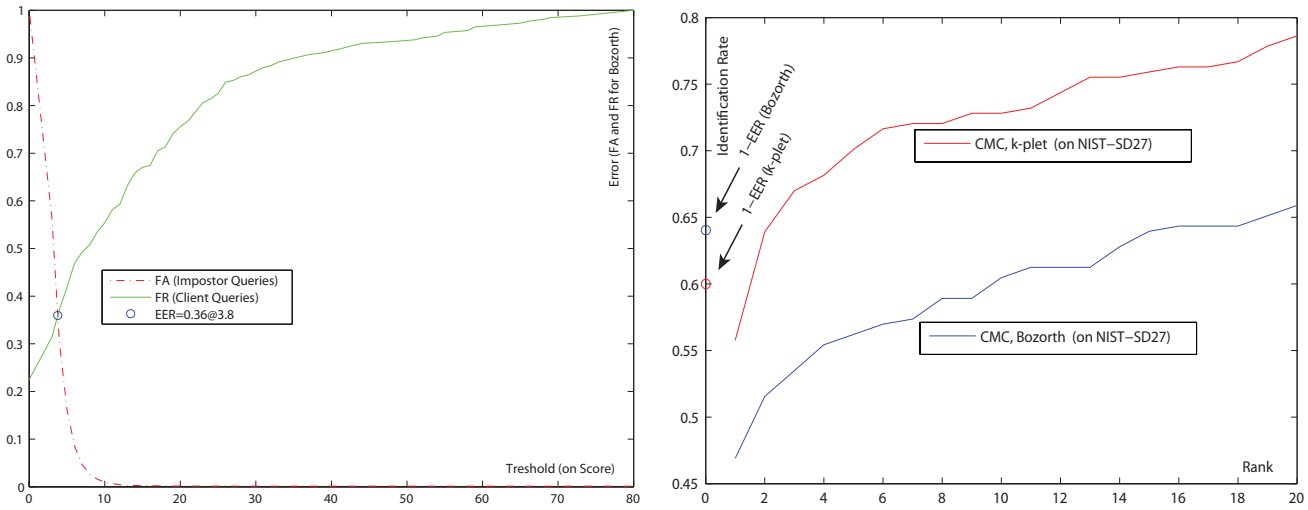

Figure 2. The FA and FR performance of the Bozorth3 and the CMC curves