A Fingerprint Matching Model using Unsupervised Learning Approach

Nasser S. Abouzakhar and Muhammed Bello Abdulazeez

School of Computer Science, The University of Hertfordshire,
College Lane, Hatfield AL 10 9AB, Hertfordshire, UK
{N.Abouzakhar, M.B.Abdulazeez}@herts.ac.uk

Abstract

The increase in the number of interconnected information systems and networks to the Internet has led to an increase in different security threats and violations such as unauthorised remote access. The existing network technologies and communication protocols are not well designed to deal with such problems. The recent explosive development in the Internet allowed unwelcomed visitors to gain access to private information and various resources such as financial institutions, hospitals, airports ... etc. Those resources comprise critical-mission systems and information which rely on certain techniques to achieve effective security. With the increasing use of IT technologies for managing information, there is a need for stronger authentication mechanisms such as biometrics which is expected to take over many of traditional authentication and identification solutions. Providing appropriate authentication and identification mechanisms such as biometrics not only ensures that the right users have access to resources and giving them the right privileges, but enables cybercrime forensics specialists to gather useful evidence whenever needed. Also, critical-mission resources and applications require mechanisms to detect when legitimate users try to misuse their privileges; certainly biometrics helps to provide such services. This paper investigates the field of biometrics as one of the recent developed mechanisms for user authentication and evidence gathering despite its limitations. A biometric-based solution model is proposed using various statistical-based unsupervised learning approaches for fingerprint matching. The proposed matching algorithm is based on three various similarity measures, Cosine similarity measure, Manhattan distance measure and Chebyshev distance measure. In this paper, we introduce a model which uses those similarity measures to compute a fingerprint's matching factor. The calculated matching factor is based on a certain threshold value which could be used by a forensic specialist for deciding whether a suspicious user is actually the person who claims to be or not. A freely available fingerprint biometric SDK has been used to develop and implement the suggested algorithm. The major findings of the experiments showed promising and interesting results in terms of the performance of all the proposed similarity measures.

1.0 Introduction

The growing dependence of modern society on information and communication technologies has become inevitable. Due to the recent explosive boom in the field of communications and transportation, intelligent systems provide access control to various resources such as information, financial data/institutions, hospitals, airports, countries and so on. Providing appropriate authentication mechanism not only ensures that the right users have access to resources but gives legitimate users the right privileges. Also, these resources need mechanisms to detect when invalid users try to misuse their privileges. Certainly biometrics helps to provide such services. Because of the nature of those resources and their reliance on computer systems to achieve effective security there is an increased need for stronger authentication mechanisms. To authenticate a user a computer system can use one of the three authentication methods [1]:

- Knowledge based, i.e. something you know e.g. password.
- Token based, i.e. something you have, e.g. token.
- Biometric based, i.e. something you are, e.g. a measurable trait.

Systems can combine one or more approaches of the same method [2]. Also systems combine one or more methods [3]. All these are done to achieve high level of security in systems. Any approach is chosen based on the requirements of the underlying system. In recent years there has been a surge in the use of biometrics for human authentication [4]. Because biometrics can be used for human authentication and hence access control, it provides several advantages as compared to other authentication mechanisms. Biometrics could reduce the likelihood that an attacker can present an identifier to gain unauthorised access. However, biometrics is also not perfect, as it has its own vulnerabilities. The biometric system by itself has different modules and each module can be vulnerable to some form of attack. Again, each individual biometric (e.g. fingerprint, iris, face, voice, hand geometry and so on) has its own limitations. So, these are some of such issues that will be looking at in this paper.

Due to the complex nature of biometric systems they have been of interest to a variety of seemingly unrelated disciplines. These disciplines include computer security, image processing, pattern recognition, mathematics, and so on. The authors will be looking at how well biometric systems will perform when a fingerprint matching algorithm is implemented using different unsupervised learning-based similarity measures. Our work investigates three unsupervised learning approaches which assume no prior knowledge about what could be the matching fingerprint. In recent years, fingerprint is one of the most widely used biometrics and it has strong user acceptance [5]. Figure 1 shows a high level model of our solution as it indicates the task of matching a user's fingerprint with a number of stored templates. The task model is scanning through a database of fingerprints to identify any match.

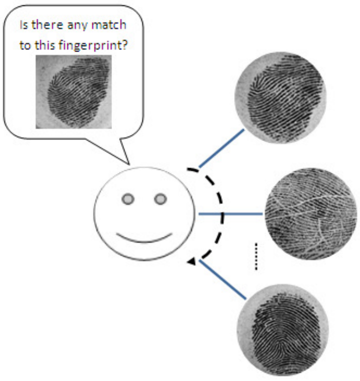


Figure 1: The task model

A fingerprint is a pattern of ridges and valleys on the surface of the fingertip. Fingerprints of identical twins are different and also prints of different fingers of individual are different. The accuracy of currently available fingerprint recognition systems is adequate for authenticating few hundreds of users but as the number of users increase the accuracy decreases, this makes the deployment of fingerprint-based authentication in large systems a problem [5]. The solution to this problem is to provide prints of multiple fingers of an individual. Another problem is that they require large amount of computational resources especially when they are in identifying mode. Also fingerprints of some people may not be identifiable due to genetic factors, aging, environmental or occupational factors (the hands of a construction worker can have many cuts and bruises).

It is possible to spoof a fingerprint, either by physically cutting the finger or making a fake fingerprint [6] [7]. Developing a fake fingerprint could be achieved using artifacts left on a scanning device. Also, trace of the prints used to authenticate legitimate users can be used to fool the access control system.

2.0 Model and Assumptions

Biometric systems attempt to provide a reliable access to secured systems and/or buildings using what a person is rather than what s/he has (ID, password, ATM cards). In this section we introduce a model which uses a statistical-based similarity measure to compute a fingerprint’s matching factor. The calculated matching factor is based on a certain threshold value which will be used for deciding whether the user is actually the person who claims to be or not.

Figure 2 shows the proposed fingerprint matching model, including the image sensing to input fingerprints, feature extraction, template generation, templates database and finally the matching process as follows:

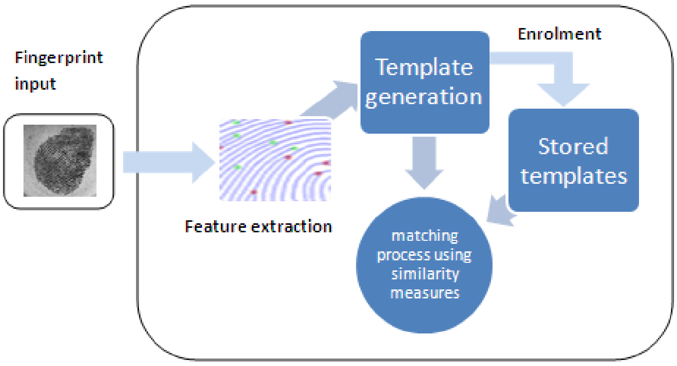


Figure 2: The proposed solution model

- Sensor: The function of the sensor is to capture the fingerprint image using a scanner. The model shows a raw fingerprint image capture using optical scanner. This stage could be followed by a pre-processing stage to improve the quality of the captured data, e.g. by improving the image quality through increasing brightness.
- Feature extraction: This is the process of extracting the relevant features which will be used for comparison purposes. This is a critical phase in the model because if a wrong feature is extracted that would have a negative impact on the final decision. In this stage all unnecessary information is discarded and the minutiae or bifurcations (a ridge splitting into two) and ridge ending points are recognized [8]. The fingerprint image is then divided into single pixel units, as shown in figure 3. The bifurcations are red-coloured points while the ridge endings are green-coloured points, as shown in figure 3.

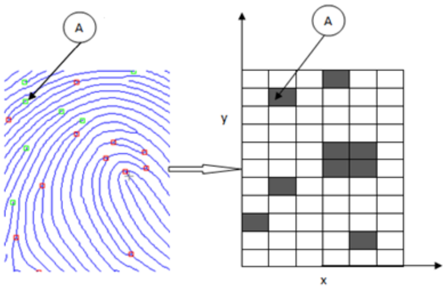
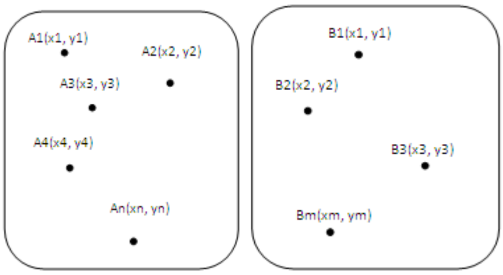


Figure 3: Feature extraction process

- Template generator: A template is the numerical representation of the fingerprint image. All the pixel units are used to represent the extracted feature points in a form of a (x, y) Cartesian coordinates.

For example, the feature point A is converted from a point on an image to a point on the Cartesian plane in a form of (x, y) coordinates. All other feature points were allocated to their nearest (x, y) points. If two or more feature points were close to each other, they can be represented by one single pixel unit i.e. a single (x, y) point. Figure 4 shows an example of allocated (x, y) points to two different fingerprints. These specific fingerprint points should be originally identified during the feature extraction stage.



(a) (b)

Figure 4: A representation of two different fingerprints

- Stored Templates: In this model, *enrolment* is the process of recording a template into the system for further authentication purposes [9]. These templates are stored for future authentication purposes. The fingerprint templates/images used for the model solution are from the Fingerprint Verification Competition (FVC2006) database. This database used images capture for the FVC2006 competition purposes. The database was designed to test the proposed fingerprint identification models to their limit. The stored templates comprise images from four different databases as follows:

    - Database 1: low-cost optical sensor “Secure Desktop Scanning” by KeyTronic
    - Database 2: low-cost capacitive “TouchChip” by ST Microelectronics
    - Database 3: optical sensor “DF-90” by Identicator Technology
    - Database 4: synthetic fingerprint generation

All the databases consist of fingerprints with the same resolution of 500dpi. We used all those four different databases to develop our solution model.

- Matching process: The matching process compares two fingerprint templates and then comes up with a decision of whether the user of the system is who s/he claims to be or not. One fingerprint input will be the user's fingerprint and the other one will be taken from the stored templates or fingerprint database. So the importance of getting the matching phase right is invaluable. Part of the matching process is to compute the similarity measure between any two input templates. Three different similarity measures are used to compute the similarity between any two input fingerprints. The operation of any biometric system will depend on the performance of the similarity measure function [10]. In this stage, any decision made will be based on whether a certain threshold is reached or not.

3.0 Method

For each fingerprint whether it is the one under examination (input fingerprint) or stored in a database (templates), we characterise each fingerprint as a vector of features whose elements are defined. Then, for each vector, we calculate its dissimilarity from each other vector in the database (stored templates). Thus for each vector we produce a set of scores (as many as there are fingerprints or templates in the database, minus one). The fingerprint/template with a max score will represent the matching fingerprint/template to the input fingerprint. Following is the developed algorithm that is used for the matching process using similarity measures:

1) Initialise a variable *matchcount* to zero.
2) (a) From figure 4 (a) take the feature point A1 and compute its similarity measure with the feature point B1 in figure 4 (b). This should be a value between 1.0 and 0.0.
   (b) Store the achieved similarity measure value in a temporary array and processed to calculate the similarity measure between point A1 and B2.
   (c) Repeat (a) and (b) by considering the next feature points in finger 4 (b) i.e. B3 to B*m*.
3) Find the maximum value from 2 above and add to *matchcount*.
4) Repeat (2) and (3) above for all the other feature points in figure 4 (a) i.e. A2 to A*n*.
5) Compute the average *machcount* and get the matching percentage between the two fingerprints.

We employ three different measures of similarity, namely the cosine similarity widely used for information retrieval (IR), Manhattan (or city block) distance and chebychev distance measure, as shown in equations 1, 2 and 3 respectively.
Let the vectors *a* and *b* represent the two vectors in question. Then the cosine of the angle between the two vectors is:

$$s\left(\vec{a},\vec{b}\right)=\frac{\sum_{i=1}^{n}a_i b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}} \quad . \qquad \mathbf{(1)}$$

And the similarity between *a* and *b* is *s(a,b).* Similarly, the *city block distance* between the two vectors is simply:

$$d\left(\vec{a},\vec{b}\right)=\sum_{i=1}^{n}\left|a_i-b_i\right| \quad . \qquad \mathbf{(2)}$$

where in each case a_i is the i^{th} element of the vector *a*.

The Chebychev (similarity) distance between the two vectors *a* and *b* is the maximum distance between both vectors. The distance between ***a***=(***a****1,* ***a****2*, etc.) and ***b***=(***b****1,* ***b****2*, etc.) vectors is computed using the formula:

$$Max_i = \left|a_i - b_i\right| \quad . \qquad \mathbf{(3)}$$

where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the values of the *i*th element at vectors ***a*** and ***b***, respectively.

In some senses, the city-block distance between the two vectors is the more simple measure, as it is simply the sum of the absolute values of the differences. Furthermore, the cosine similarity measure is widely accepted in the field of IR; however, we believe that the city block distance provides a useful basis for comparison given that the application is quite different from information retrieval.
In all cases, the vectors *a* and *b* are normalised so that each feature is on the same scale. For this work, we pick a simple normalisation to z-scores. For the i^{th} element of a vector, let s_i represent some segment *s*'s score for feature *i* and μ_i represent the mean of the i^{th} feature across all fingerprint templates in the database including the input fingerprint. Similarly, let σ_i equal the standard deviation for the i^{th} feature.

Then the normalised z-score for s_i, which we have thus far called a_i, is shown in equation 4, as follows:

$$a_i = \frac{s_i - \mu_i}{\sigma_i} \; . \qquad (4)$$

4.0 Experiments and Results

Certainly, the performance of biometric systems affects the location and the context in which they are deployed. In general, there are two types of biometric systems i.e. identification and verification systems. The performance measure of these two systems differs extensively. The performance criterion for identification system is its ability to identify a biometric signature's owner [11]. However, the performance of verification systems is characterized by two errors i.e. False Acceptance Rate and False Rejection Rate (FAR and FRR). Normally the errors come in pairs, there is a FAR for every FRR value. In ideal biometric system both FRR and FAR are zero, however biometric systems are not ideal, so there is always a trade-off between these two errors. If all users are given access to the system, the FRR will be equal to zero while the FAR will be equal to one. On the other hand if all users are denied access to the system, the FRR will be equal to one and the FAR will be zero. Certainly a parameter is used (or adjusted) to obtain the desire error rates. This parameter is the Decision threshold; a decision threshold is a limit that decides whether a matching value is greater than the limit value. The higher the decision threshold the lower the FRR and the higher the FAR and vice-versa for lower decision threshold value. Again, these values are dependent on the requirements of the underlying (the system that the biometric protects) system. As for all biometric models, the performance of this model can be affected by the following factors [4].

1) Quality of biometric input and enrolment data
2) The characteristics of the underlying feature extraction, and
3) Matching algorithm

In our experiments, 80 sample fingerprints are provided. Each input fingerprint is compared with all the eighty fingerprint images using the three different matching algorithm suggested. The decision threshold is altered in order to see how well these matching algorithms perform under various security settings. The FVC2006 database contains 80 different fingerprint images for each user as the fingerprint images are from ten fingers. Each user provides 8 different impressions of the same finger (8 × 10 = 80).

To do the test, an image is taken at random from any of the 8 impressions of a finger. The image is then compared with all other images in the database (including itself and other impressions of the same fingerprint). The comparison is done using all three different matching methods. The system returns a matching score of between 0 and 100 between the chosen fingerprint image and all other images. Note that the original values obtained are in the range between 0 and 1, but are converted to percentage here for better representation. It should be noted that for all the three matching algorithm the system returns 100 matching score for itself. This is done for all the 10 fingerprint sets.

Ideally, when a decision threshold is chosen (e.g. 90), all eight fingerprint images in the set that the base image (input fingerprint) belong to should return a value greater than or equal to 90 and all other 72 images should return a value less than 90. This evaluation will attempt to compare the performance of the proposed algorithms based of the previously mentioned characteristics of biometric systems. As the decision threshold (T) represents the minimum matching value above which a fingerprint image is considered as a match, the two main factors that are used to evaluate the performance of the proposed model are:

False Match Rate (FMR): is the percentage of fingerprint images that have value greater than or equals to T but are not in the base fingerprint set.

False Non-Match Rate (FNMR): is the percentage of fingerprint images that have values less than T and are in the base fingerprint set.

The FMR and FNMR of the algorithms are calculated by taking one impression of fingerprint image from each of the 10 fingerprint sets and comparing it with all the fingerprints in the database (80 prints). When the comparison is done FMR and FNMR are calculated as follows:

$$\text{FMR} = \text{number of images that falsely match}/72\ . \qquad \textbf{(5)}$$

$$\text{FNMR} = \text{number of } \textbf{true} \text{ images that did not match}/8\ . \qquad \textbf{(6)}$$

.

For equation 5 the desired value is 0, therefore if 60 images falsely match then FMR = 60/72 = 0.83, while if only 2 images match then FMR = 2/72 = 0.03. Note that the number 72 is derived from equation 7, as follows:

$$\text{total - total number of true images} = 80 - 8 = 72\ . \qquad \textbf{(7)}$$

For equation 6 the desired value is 0, therefore if 6 true images do not match then FNMR = 6/8 = 0.75, while if only 1 true image did not match then FNMR = 1/8 = 0.13. Note that the number 8 represents the number of impressions of the same finger (number of true images).

After obtaining the values for all the fingerprints, the average value is calculated and is taken as the FMR/FNMR for that particular algorithm. Figures 5 and 6 show the achieved FMRs and FNMRs for all the three similarity measures when T = 90 and T = 80 respectively. Both figures show that Manhattan (or city block) and Chebyshev have outperformed the cosine similarity measure in terms of FMR but vise versa in the case of FNMR. Cosine similarity has a very high FMR (63% when T = 90 and 73% when T = 80). However, Manhattan distance measure has FMR = 26 and FNMR = 50 for T = 80. When T = 90 the FMR is at an impressive rate of 10% but the FNMR is disappointingly high 66%. The Chebyshev distance performs similar to its Manhattan compatriot with impressive low FMR but a relatively poor FNMR especially for T=90.

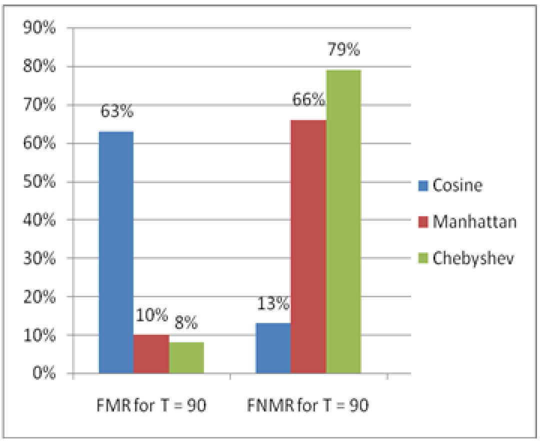


Figure 5: FMR and FNMR for all the three similarity measures (T = 90)

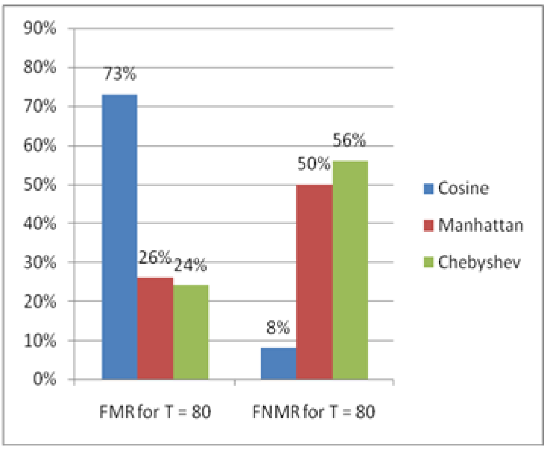


Figure 6: FMR and FNMR for all the three similarity measures (T = 80)

The Chebyshev distance measure achieved the minimum FMR of 8% for T = 90. However, Cosine similarity achieved the minimum FNMR of 8% for T = 80. The results show the effect of altering the decision threshold T. In general, when the value of T decreases the value of FMR increases however, the value of FNMR decreases. This phenomenon is proven to be true for all the distance measures. For example, when the value of T decreases from 90 to 80 the Manhattan's FMR increased from 10% to 26% and the value of FNMR decreased from 66% to 50%.

5.0 Conclusion

This paper introduced the problem of identifying fingerprints by automatically extracting specific features from the input fingerprint and evaluating those features for patterns of consistent information. In order to achieve that, we defined specific fingerprint features that characterise certain fingerprint properties. It is important to know where and how the system will be deployed in order to perform the evaluation in a better context. In view of this, we realised that biometric systems have two different modes of operation (Identification and Verification) and the performance measures of these modes vary extensively. Three distance measures approaches were used to develop our solution model, the cosine similarity (distance) measure, Manhattan (or city block) distance and chebychev distance measures have been used to measure the distance of each input fingerprint from every other fingerprint in the Templates database. Our experiments showed encouraging results and our research indicated a significant unsupervised learning power in the application of biometric security.

By means of evaluation, as well as empirical evidence, we are able to determine the effectiveness of the developed models and assumptions. The performance of the three developed identification models has been evaluated and the results indicate that both distance measures Manhattan and chebychev outperformed cosine similarity measure in terms of FMR but vise versa in the case of FNMR. However, all models have achieved a significant increase in the matching rates in terms of identifying fingerprints.

References

1 Dhamija R, Perrig A, Deja Vu, A User Study Using Images for Authentication. In: 9th Conference on USENIX Security Symposium, vol 9, Springer, California, 2000

2 Toh K, Xiong W, Yau W, Jiang X, Combining fingerprint and hand-Geometry Verification Decisions. In: AVBPA 2003. LNCS. 2688, pp. 688-69. Springer, Berlin, 2003

3 Zheng Y, Xia J, He D, Trusted User Authentication Scheme Combining Password with fingerprint for mobile devices. In: IEEE- International Symposium on Biometrics & security technologies, ISBAST, Islamabad 2008

4 Ratha N, Connell J, Bolle R, (2001). Enhancing Security and Privacy in Biometric-Based Authentication Systems. J. IBM Systems. vol 40 no. 3, 614--634

5 Jain A K, Ross A, Pankanti S, Biometrics(2006). A Tool for Information Security. IEEE Transactions on Information Forensics And Security. vol. 1 no. 2

6 Matsumoto T, Matsumoto H, Yamada K, Hoshino S, Impact of Artificial 'Gummy' Fingers on Fingerprint Systems. In: Proceedings of International Soc. Optical Eng. (SPIE), vol. 4677 (2002)

7 Sean P, Booz A, Harry W, Fingerprint Readers: Vulnerabilities to Front- and Back-end Attack, California, McGraw Hill, (2007)

8 Johnson S, Fingerprint Software Development Kit (SDK). version 1.3. (2005)

9 Reid P, Biometric: for Network Security. New Jersey, Prentice Hall (2004)

10 Roman Y, Venu G, Similarity Measure Functions for Strategy-Based Biometrics. In: Proceedings of the World Academy of Science and Technology, vol 18 (2006)

11 Jonathan P, Martin A, Wilson C, Mark P, An Introduction to Evaluating Biometric Systems, In: IEEE Computer Society Press, Los Alamitos, CA (2000)