A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Evaluation on Biometric Accuracy Estimation Using Generalized Pareto (GP) Distribution

Shigefumi Yamada[1], Tomoaki Matsunami[2]

Abstract: The accuracy of biometric authentication technology is becoming more sophisticated with its progress. For this reason, a huge number of biometric samples are required for accuracy evaluation, and the increased collection cost is an issue for biometric vendors. This work establishes a biometric accuracy estimation method using an extreme value theory to reduce the collection cost. It also explains the estimation procedure of false match rate using the generalized Pareto distribution and shows results applied to the face, gait, and voice comparison score data with an estimation effect of about 5–10 times. We investigate the criteria for the applicability of extremum statistics through application cases.

Keywords: accuracy evaluation, false match rate, rule of three, extreme value theory, generalized Pareto distribution.

1 Introduction

As the utilization of Information Technology permeates various corporate activities and consumer life, high-precision biometric authentication (face, fingerprint, vein, etc.) is required to effectively provide and use more secure, safe, and detailed services. Iris, voice, signature, and walk are being used for personal authentication.

The method for evaluating the accuracy of biometric authentication is specified and recommended in the ISO / IEC19795 (Biometric performance testing and reporting) [I21] series and has been applied to many biometric authentication devices and software. On the other hand, with the progress and higher accuracy of biometric authentication technology, a huge number of biometric samples (data obtained from fingerprints, faces, veins, etc.) are required for the evaluation. There is false match rate (FMR) and false non match rate (FNMR) as biometric recognition accuracy. Traditionally, Rule of 3 have been used to estimate the number of data required to determine FMR with a 95% confidence interval. The Rule of 3 requires several comparisons that is three times the reciprocal of the error rate you want to evaluate. Each time a biometric vendor develops a product, the accuracy must be verified with many biometric samples. The accuracy of

[1] Japan Automatic Identification Systems Association, Biometrics Research Group, Performance testing SIG., FK Bldg.7F, 1-9-5, Iwamoto-cho, Chiyoda-Ku, Tokyo, 101-0032, Japan, yamada.shige@fujitsu.com
[2] Japan Automatic Identification Systems Association, Biometrics Research Group, Performance testing SIG., FK Bldg.7F, 1-9-5, Iwamoto-cho, Chiyoda-Ku, Tokyo, 101-0032, Japan, t.matsunami@fujitsu.com

current biometric products is extremely high. For example, the number of biometric data required to evaluate FMR is shown in Tab. 1. Therefore, the cost of collecting biometric samples is a burden for biometric vendors.

FMR	Required number of non-mated trials in the case of zero false matches	Required number of test subjects in the case of full cross-comparison
0.001%	Over 300,000	Over 775
0.0001%	Over 3 million	Over 2450
0.00001%	Over 30 million	Over 7746

Tab. 1: Number of biometric samples required to evaluate FMR

If the tail of the non-mated comparison score distribution that is unstable due to the small number of scores could be modeled appropriately, the score distribution could be extrapolated and a more accurate FMR could be estimated with the small numbers. Therefore, this work adopted the idea of extreme value theory (EVT), which is a statistical inference about the probability of occurrence of rare events, and used the EVT [C01] as a distribution model of the region where false match occur. An attempt to solve this problem is made by adopting the generalized Pareto (GP) distribution.

There are a few cases in which the EVT has been applied to the accuracy evaluation of biometrics. For the design of large-scale identification systems, it has been proposed to approximate the tail of the comparison score distribution for others with the GP distribution [HJ05]. At the tail of the distribution, which varies widely due to the lack of data, the GP distribution provided a more reliable estimate of the FMR than the measured values. When determining the decision threshold for the 1:1 comparison, it is important to appropriately approximate the tail of the comparison score distribution between the mated and non-mated pairs. As a method of estimating the tail of the distribution of comparison scores, the application of GP, which is one of the extreme value statistics, has been proposed [ZFJV08]. Extreme statistics are used for score normalization in the score level fusion of a multimodal biometric system [WART10]. The generalized extreme value (GEV) distribution, especially the Weibull distribution, of the extreme value statistics is applied to estimate the tail of the comparison score distribution of the non-mated and mated pairs. Robustly estimating the tail of the distribution, which was unstable due to the small comparison score, improves the accuracy of the score fusion. Similarly, the GEV distribution has been applied for score normalization in the score level fusion, and better discrimination results have been reported compared to those of conventional Z-score normalization [RSP15].

With these prior arts, the FMR can be estimated with high reliability by approximating the tail of the comparison score distribution with extremum statistics. Furthermore, the FMR that could not be measured with actual data can be estimated by extrapolating the tail of the comparison score distribution. Using this property, it can be expected that the desired FMR can be estimated with the smaller number of samples than using the rule of three. On the other hand, the applicability of the EVT for biometric authentication has

not been discussed much in the prior arts. In applying the EVT, the comparison scores should be independent and identically distributed (i.i.d.). As a unique problem of biometric technologies, the distribution of comparison scores for non-mated pairs is known to be biased. For example, there are genetically identical/similar samples such as twins. The authenticated user who achieves a high score for many registered users is called "Wolf," and the registered user who gives a high score to a large number of authenticated users is called "Lamb." Such users can easily cause many false acceptance errors and affect the i.i.d. condition due to the high comparison scores. To spread the application of the EVT in the accuracy evaluation of biometric technology, it is important to accumulate applicable and difficult-to-apply cases through various application cases and clarify the criteria for applicability.

The main contributions of this work are as follows:

- The GP distribution is applied to various comparison score data of face, gait, and voice, and the estimation effect of the FMR is evaluated through an experiment that estimates the FMR required for evaluation from a small amount of data.
- The criteria for the applicability of extremum statistics are investigated through application cases.

This work is based on the research results of the project conducted by the Ministry of Economy, Trade and Industry in Japan from FY2019 to FY2021 [ME20].

2 Extreme Value Theory (EVT)

EVT [C01] typically targets natural phenomena that would cause very large disasters, such as heavy rains, large earthquakes, high waves, typhoons, and droughts, and their potential. It has been used for scale prediction and evaluation, i.e., in EVT, statistics are focused on the extreme value data at the tail, not the main part of the population distribution, which is noted in general statistical applications. It is characterized by making estimates based on target extrapolation.

In the method described in this work, extreme value data is classified into three types: "maximum data in a large block," "upper r data in that block," and "all data exceeding sufficiently large values in the observed data." Distributions that apply to these data include the GEV distribution, the simultaneous asymptotic distribution of the top r ordinal statistics (rGEV), and the GP distribution. These distributions assume that the original data are i.i.d. and nondegenerate.

Because the GP distribution is used in this work, its definition is described. The threshold excess data $\{x_1, x_2, \ldots, x_n\}$ are measured values of random variables that follow the generalized Pareto distribution $GP(\sigma, \xi)$ independently and identically. The GP distribution [C01] has a cumulative distribution function:

$$F(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - exp\left(-\frac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases} . \quad (1)$$

where σ is the scale parameter for the GP distribution, ξ is the shape parameter, and μ is the location parameter (threshold).

The advantage of the EVT is that most continuous distributions, such as the normal and exponential distributions, belong to the suction region of one of the extreme value distributions. Therefore, the tail region can be approximated by the GEV or GP distributions in many continuous distributions.

3 Biometric Accuracy Estimation Method Using GP distribution

If the tail of the non-mated similarity score distribution that was unstable due to the small number of scores can be accurately approximated using the GP distribution, then the FMR that could not be measured due to the lack of scores can be estimated using extrapolating the score distribution. This method's procedure is described below.

1. Extraction of extreme value data

By selecting the threshold value μ, the data exceeding μ is extracted as extreme value data. When selecting μ, the parameter estimation of the scale parameter σ and the shape parameter ξ is performed while changing μ. μ is plotted on the x-axis, while the estimated value is plotted on the y-axis. If the estimated values can be constant to the right of a certain value, their minimum value can be determined as a threshold.

2. Estimation of the GP distribution parameters

Assuming that the distribution of the extreme value data follows the GP distribution, the scale and shape parameters σ and ξ, respectively, of the GP distribution function are obtained using the maximum likelihood estimation method.

3. Diagnosis of the estimation results

Create a quantile–quantile plot (Q–Q plot) with the percentile values of both the GP distributions obtained via the parameter estimation and actual measurement values used for the estimation on the x- and y-axes, respectively. The Q–Q plot is a method for graphically comparing two probability distributions. If the two distributions to be compared are similar, then the points in the Q–Q plot are located near the straight line y = x. In the observation of the Q–Q plot, the suitability of the model shall be confirmed by the degree of deviation from the straight line y = x. Here there is no method for quantitatively determining the degree of deviation, and the suitability of EVT is conventionally determined via human eyes [C01].

4. Estimation of the FMR

Based on the probability density function of the obtained estimation model, the FMR when the threshold is set can be obtained. Fig. 1 shows an example of a graph plotting the FMR estimated by the GP distribution. The y-axis is plotted on the common logarithm axis to improve visibility. The solid blue line shows the measured value, and the dashed-red line shows the estimated FMR by the GP distribution. The green vertical line indicates the maximum value of the measured value and the score value larger than that indicates the extrapolated value. The FMR at any threshold can be obtained from this graph. If the threshold is selected as 50, the GP estimate is $10^{-7.6224} = 2.3858 \times 10^{-8}$.

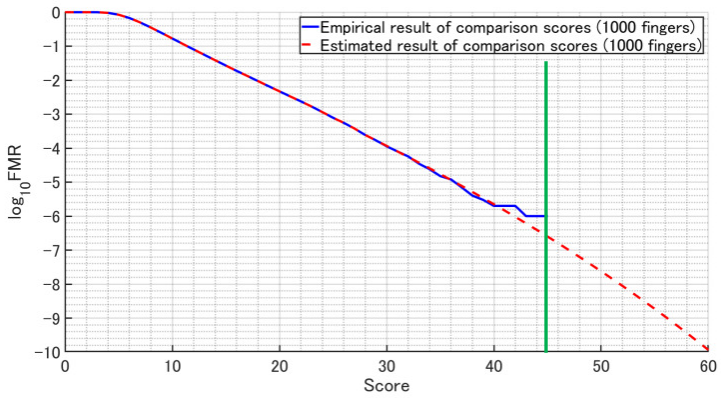


Fig. 1: Estimation result of FMR using GP distribution

4 Experimental Results and Discussions

4.1 Database and Implementation Details

To investigate the effectiveness of the GP distribution in biometric accuracy evaluation, we tried to apply the GP distribution to non-mated similarity score data in face, gait, and voice recognition. Tab.2 shows the details of the similarity score data used in these experiments. As a protocol, the test data was created by randomly extracting a small number of data (10% or 20%) from all score data. GP was applied to these test data, and the validity of the estimation result was verified by the Q–Q plot. If there were no problems in the verification results, the FMR was calculated. The calculated FMR was compared to the FMR measured from all score data. If the results were close, we could, in a sense, estimate the FMR that would require 5 to 10 times more scores. When the approximation using the GP distribution was not appropriate for 10% of the test data, the evaluation was made by increasing the number of scores to 20%.

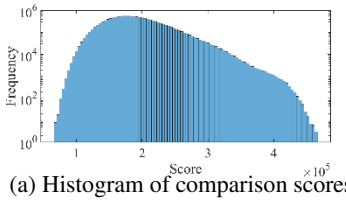
Biometrics	Corpus of biometric samples	Comparison score dataset	Number of comparisons
Face	FRGC	Idiap BIOSCOTE 2014, Face Recognition Grand Challenge v2.0 [L14]	4 million
Gait	GEI	Osaka University, Gait Energy Image (GEI) [Ha12], [Ng12]	13.73 million
Voice	SRE2012	Idiap BIOSCOTE 2014, NIST Speaker Recognition Evaluation 2012 [L14]	13.49 million

Tab. 2: Description of the non-mated similarity score data used in the experiments

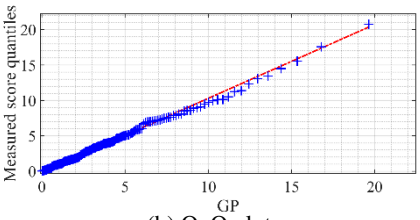
4.2 Results

The face recognition grand challenge (FRGC) and gait energy image (GEI) showed good results for the GP distribution. Fig. 2 (a) shows a histogram of the similarity score of the FRGC. Twenty percent of the scores (approximately 800,000 scores) were randomly extracted from all score data and applied to the GP distribution as test data. Fig. 2 (b) shows the Q–Q plot when approximated by the GP distribution (threshold value μ = 0.485), while Fig. 2 (c) shows the comparison between the estimated and measured FMRs. The Q–Q plot is located near $y = x$, and the right tail of the similarity score distribution can be well approximated by the GP distribution. With the FRGC score, the FMR for all scores can be estimated from 20% of the scores, so it can be said that the FMR requiring about 5 times more data can be estimated by the GP distribution. Finally, Figs. 2 (d), (e), and (f) show the experimental results of the GEI. Ten percent of comparison scores (about 1.37 million scores) were randomly extracted. The threshold μ was 79.5. The Q–Q plot is located near $y = x$. With the GEI scores, the FMR that requires about 10 times more data with the GP distribution can be estimated. In Fig. 2 (f), the red line is slightly below the blue line because the subset randomly extracted for testing is below the entire set.

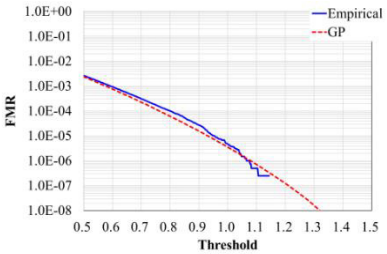
The result of SRE2012 in Fig. 3 is shown as an example of the difficult application of the GP distribution. Twenty percent of the similarity scores (about 2.68 million scores) were randomly extracted. The threshold μ was 2.8. In Fig. 3 (a), the score values are in a very high region and the right tail of the similarity score distribution is discretely distributed. In Fig. 3 (b), the upper right of the Q–Q plot deviates from the line $y = x$, indicating that the GP distribution cannot adequately approximate the right tail of the similarity score distribution. Moreover, the estimated FMR was different from the measured FMR in Fig. 3 (c). It can be seen that the application of the GP distribution is difficult when its right tail is discrete.



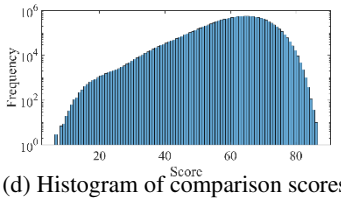
(a) Histogram of comparison scores



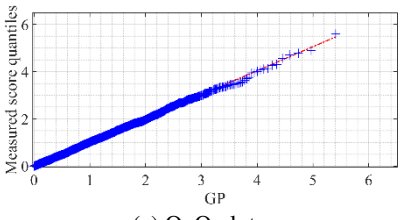
(b) Q–Q plot



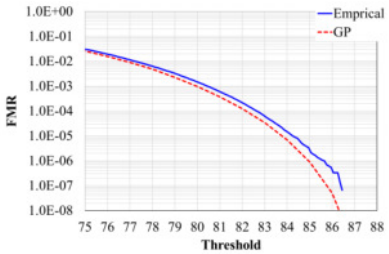
(c) FMR between measured value and GP



(d) Histogram of comparison scores

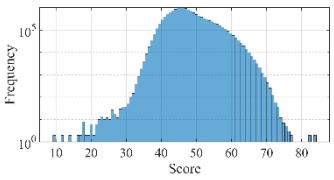


(e) Q–Q plot

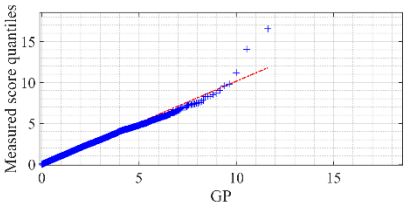


(f) FMR between measured value and GP

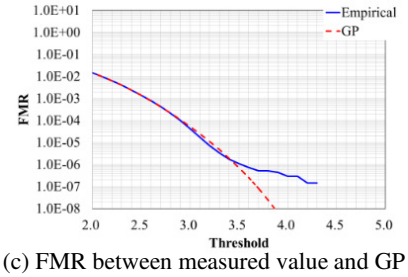
Fig. 2: Results of FRGC ((a), (b), (c)) and GEI ((d), (e), (f)) using GP distribution



(a) Histogram of comparison scores



(b) Q–Q plot



(c) FMR between measured value and GP

Fig. 3: Results of SRE2012 ((a), (b), (c)) using GP distribution

5 Conclusions and Future Works

This work shows the procedure for applying EVT, especially the GP distribution, to the accuracy evaluation for biometrics. This method is applied to the non-mated similarity score data of face, gait, and voice recognition, and the estimation effect is confirmed, which is 5 times for FRGC and 10 times for GEI. It was confirmed that it is difficult to apply the GP distribution in SRE2012 because there are discrete scores at the tail of the similarity score distribution. In the future, we plan to apply this method to the various cases of biometric technology and try conditions such as data size to clarify the applicable conditions of this method. In addition, we would like to make it easier to use this method by a creating criteria for determining the suitability of the GP distribution by human eyes on the Q–Q plot.

References

[I21] ISO/IEC JTC 1/SC 37 Biometrics, ISO/IEC 19795-1:2021, 2021.

[C01] Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, 2001.

[HJ05] Herve J.; Jean-Christophe F.: Large-Scale Identification System Design: Chapter 9 of Biometric Systems, Springer, 2005.

[ZFJV08] Zhixin, S.; Frederick, K.; John, S.; Venu, G.: Modeling Biometric Systems Using the General Pareto: In proceedings of SPIE, March 2008.

[WART10] Walter, S.; Anderson, R.; Ross, M.; Terrance, B.: Robust Fusion: Extreme Value Theory for Recognition Score Normalization: ECCV 2010, pp 481-495, 2010.

[RSP15] Renu, S.; Sukhendu, D.; Padmaja, J.: Score Normalization in Multimodal Systems using Generalized Extreme Value Distribution. Conference: Proceedings of the British Machine Vision Conference (BMVC) 2014.

[ME20] Ministry of Economy, Trade and Industry, https://www.meti.go.jp/english/press/2020/0923_003.html, 9.23.2020

[L14] Laurent, S.: Scalable Probabilistic Models for Face and Speaker Recognition, PhD thesis, 2014. http://publications.idiap.ch/index.php/publications/show/2830

[Ha12] Haruyuki, I.; Mayu, O.; Yasushi, M.; Yasushi, Y.: The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition: IEEE Trans. on Information Forensics and Security, Vol. 7, No. 5, pp. 1511-1521, Oct., 2012. (Data Set 1, 2, and 4).

[Ng12] Ngo, T.; Yasushi, M.; Hajime, N.; Yasuhiro, M.; Yasushi, Y.: Performance Evaluation of Gait Recognition using the Largest Inertial Sensor-based Gait Database: Proc. of the 5th IAPR Int. Conf. on Biometrics, Paper ID 182, pp. 1-7, New Delhi, India, Mar., 2012.(Date Set 3).