IET Biometrics

Research Article

Searching for doppelgängers: assessing the universality of the IrisCode impostors distribution

ISSN 2047-4938
Received on 27th August 2015
Revised on 26th October 2015
Accepted on 27th October 2015
doi: 10.1049/iet-bmt.2015.0071
www.ietdl.org

John Daugman ✉, *Cathryn Downing*

Faculty of Computer Science and Technology, University of Cambridge, Cambridge, UK
✉ *E-mail: john.daugman@cl.cam.ac.uk*

Abstract: The authors generated 316,250 entire distributions of IrisCode impostor scores, each distribution obtained by comparing one iris against hundreds of thousands of others in a database including persons spanning 152 nationalities. Altogether 100 billion iris comparisons were performed in this study. The purpose was to evaluate whether, in the tradition of Doddington's Zoo, some individuals are inherently more prone than most to generate iris false matches, while others are inherently less prone. With the standard score normalisation disabled, a detailed inter-quantile analysis showed that meaningful deviations from a universal impostors distribution occur only for individual distributions that are highly extreme in both their mean and their standard deviation, and which appear to make up <1% of the population. In general, when different persons are compared, the IrisCode produces relatively constant dissimilarity distances having an invariant narrow distribution, thanks to the large entropy which lies at the heart of this biometric modality. The authors discuss the implications of these findings and their caveats for various search strategies, including '1-to-first' and '1-to-many' iris matching.

1 Introduction

The well-known 'birthday problem' asks how large a group of people must be assembled, chosen randomly, before it becomes more likely than not that at least one pair of them have the same birthday. It is easy to calculate that this occurs once there are at least $N = 23$ persons. There exists an analogous 'biometric birthday problem': for a given similarity threshold yielding some specified false match rate (FMR) for single comparisons, how many different persons must a database contain before it becomes likelier than not that there is at least one biometric collision? Weak biometric technologies such as face recognition are usually tested and operated at the very undemanding criterion of FMR = 0.001, which means that any given pair of random persons have probability 0.999 of *not* being matched to each other. Since N persons make $N(N–1)/2$ possible pairings, a biometric collision becomes likelier than not when $(0.999)^{N(N-1)/2} < 0.5$ and this occurs when there are just $N = 38$ or more persons. Consider, for example, the picture gallery of readily confusable yet unrelated persons presented as 'doppelgänger' pairs in [1]. Indeed, to a human observer, the doppelgängers may even appear more similar to each other than individuals typically resemble *themselves* after changes of pose, expression, illumination geometry, or age. This paper investigates: (i) why, in contrast, there is a safe and fairly constant dissimilarity distance whenever different irises are compared, a property obviously beneficial for biometric collision avoidance; (ii) whether the relatively narrow distribution of such scores has a universal form; (iii) the limits to the invariance of this distribution; (iv) implications for search strategies, including '1-to-many' and '1-to-first'; and (v) whether any evidence can be found for iris doppelgängers.

These questions are important because an invariant impostors distribution is, or would be, highly advantageous for any biometric modality. It means that a given dissimilarity score threshold can be immediately translated into a false match probability and a confidence level, calculated as the cumulative below that score threshold under the universal impostors probability distribution for that modality, regardless of who generates the scores. It also allows straightforward extrapolation from a single-comparison False Match probability given some score, to the *net* error probability if the score was observed only after searching a large database that may be of national scale (as is now done daily in India with enrolment underway of all 1.2 billion citizens [2] within 3 years). Thus, the *number* of alternative iris comparisons that are made, before a given best match is encountered, can be taken into account in its interpretation. Finally, if a given biometric modality *cannot* assume a universal impostors distribution, then any observed similarity score must be further qualified by whether the subject is the type of person who has many doppelgängers, or few.

2 Methods and database

We generated 316,250 entire distributions of 'impostor' (i.e. different eye) iris dissimilarity scores for detailed inter-quantile analysis and comparison, using a large database including persons of 152 nationalities. Each of these distributions was obtained by comparing one eye against a gallery of several hundred thousand others. Statistics harvested from these distributions enabled us to order them according to distributional properties and to perform inter-quantile analyses, answering questions such as: 'How different are the impostors distributions whose means (or whose standard deviations, henceforth 'stnd-devs') are in the uppermost 99.99th percentile when compared with those in the lowest 0.01th percentile? In such metrics, how different are the top 20 distributions, and the lowest 20 distributions, from the canonical impostors distribution? How large are the effects of those differences on the FMR at a given decision threshold?' We address such questions both with the standard score normalisation, and with none.

The database, which has been described in detail already [3], was acquired by the United Arab Emirates (UAE) border-crossing security system based on iris recognition, launched in 2001 and deployed now at all 35 air, land, and sea ports of entry. Most persons who reside and work in the UAE – more than 85% – are not UAE nationals but foreign nationals. All who must apply for a visa to reside in the country are compared exhaustively against a 'negative watch-list' of persons deemed dangerous, or who have been expelled previously, or who have been denied entry for various reasons including security concerns, travelling under false documents, or work permit violations. An 'expellee' database of iris patterns were encoded as IrisCodes from persons who were expelled under an amnesty program, for the purpose of controlling re-entry. Actual images are not available, and we cannot be certain

© The Institution of Engineering and Technology 2015

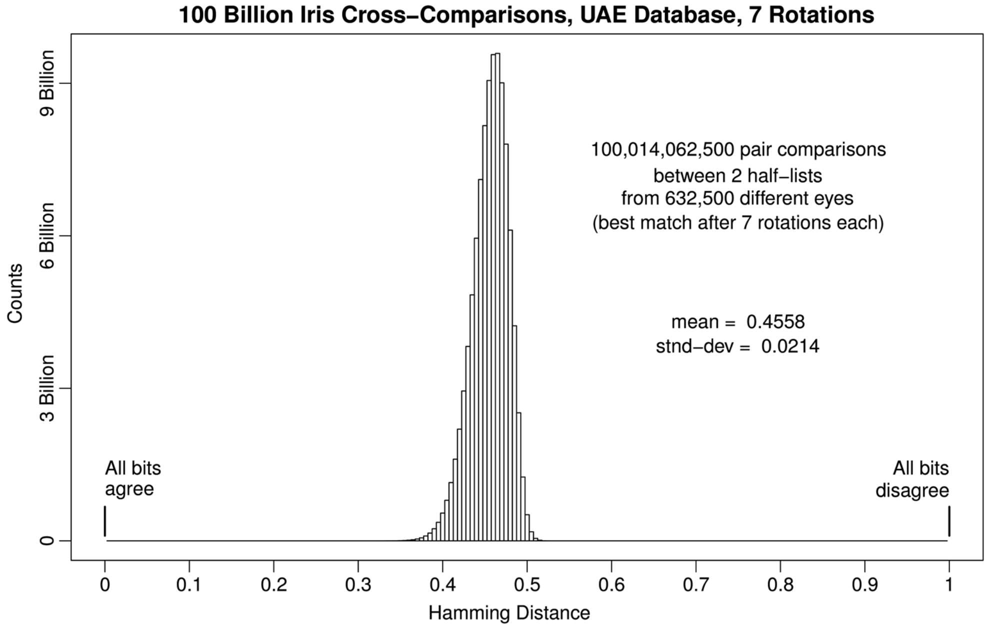

Fig. 1 *Canonical distribution of dissimilarity scores obtained by comparing each of 316,250 IrisCodes from a large Middle-Eastern database with 316,250 others, in several relative orientations to allow for unknown image tilt, recording only each best match. Hamming Distance is the fraction of bits that disagree*

how well the operational processes avoided errors. The UAE Minister of Interior, H. R. H. Sheikh Saif Bin Zayyed provided this database of 632,500 IrisCodes to the University of Cambridge, where the algorithms for iris recognition had been developed [4].

When IrisCodes from different eyes are compared, each bit pairing has equal probability of agreeing or disagreeing, because 1s and 0s are equiprobable and uncorrelated across such pairs (even between genetically identical eyes [5] such as those of monozygotic twins or one person's right and left eyes). Thus, the fraction of bits that disagree is on average 50%, and the distribution of such *fractional Hamming distances* (HDs) in simple comparisons between IrisCodes from different eyes always has a mean close to 0.5 and a relatively narrow spread around this fraction, because so many bits are compared. An example was given in Fig. 5 of [3]. However, it is necessary always to compare IrisCodes over a range of relative orientations because of unknown tilts of the head, the camera (especially if handheld), and indeed some torsional rotations of the eye around its axis. Typically, all the bits in two IrisCodes are compared in each of seven relative rotations, which amount to scrollings of the IrisCodes, with only the best match (the smallest HD) being preserved as the match score. Obviously, this selection of smallest scores creates a new 'extreme value distribution', biased towards scores smaller than 0.5 (typically around 0.45) and with asymmetric tails (negative skew). For our initial studies we generated 316,250 complete distributions of impostor scores, each distribution associated with one particular eye. We divided the 632,500 IrisCodes from the UAE database by alternation into two disjoint half-lists, one constituting the 'probes' and the other the 'gallery' (we confirmed that our results were unchanged when comparing each probe against all others in the probe list instead). IrisCode comparisons were made at each of seven rotations, and only the best match of the seven was recorded. The scores from these 100,014,062,500 comparisons between different eyes (316, 250 × 316, 250, or 100 billion for short) are all combined together in Fig. 1. The original raw distribution centred symmetrically on 0.5 (no skew) before the extreme value sampling across multiple relative tilts, comparing each of these IrisCodes with all of the others, was shown previously in Fig. 5 of [3].

3 Variation among impostors distributions

The distribution of HD scores in Fig. 1 is 99% contained within the interval $0.4 < HD < 0.5$, and so to the extent that we consider this interval to be narrow, we could say that nearly all IrisCodes computed from different eyes are roughly equidistant from each other. This is a rather striking property, and it arises from the randomness and equiprobability of bits in different IrisCodes. This is also the reason why degradation in image quality does not seem to affect the impostors distribution, unless it actually introduces some kind of coherence that *destroys* the native randomness of bits. For example, if different subjects have images acquired in the same bright local environment that produces some shared corneal reflections (e.g. bright windows or displays), this could introduce some spurious similarities in different IrisCodes and thus shift the impostors distribution leftward. In contrast, the authentics (same eye) distribution (whose mean is usually below $HD = 0.15$) can be seriously affected by optical defocus and shifted rightward; severe defocus causes the variation among pixel values to be dominated by post-optical sampling noise and thermal noise in the camera, leading to random IrisCode bits. However, introducing randomness into IrisCode bits, or increasing it, cannot degrade the impostors distribution.

Now we wish to decompose the combined distribution of 100 billion impostor scores from Fig. 1 into many sub-distributions, in order to learn whether any systematic variation may be concealed within its form. In particular, we wish to study whether individual subjects tend to have distinctive impostor match score distributions, as has been asserted generally for biometric technologies [6, 7]. We shall study this question in various ways in this paper. Initially, we use the standard [3] IrisCode matching algorithm that is used in all public deployments of iris recognition, and which incorporates an automatic score normalisation process. For this initial study we generated 316,250 entire impostor distributions using the two half-lists as described in the previous section, with each distribution being associated with one individual eye as the probe. The means of the HD scores for each of these distributions are accumulated in the blue histogram of Fig. 2, revealing an extremely narrow distribution: about three-quarters of all the mean scores fall between 0.452 and 0.457 HD. Likewise, the stnd-devs of each of those distributions of scores are accumulated in the red histogram in this figure. Once again this statistic has a remarkably narrow distribution, with about three-quarters of the stnd-devs falling between 0.020 and 0.023 HD. It appears that for iris recognition based on the IrisCode, there is very little variation in the impostors distributions for individuals. Rather, these statistics for individual distributions suggest that the histogram in Fig. 1 is quite a universal impostors

© The Institution of Engineering and Technology 2015

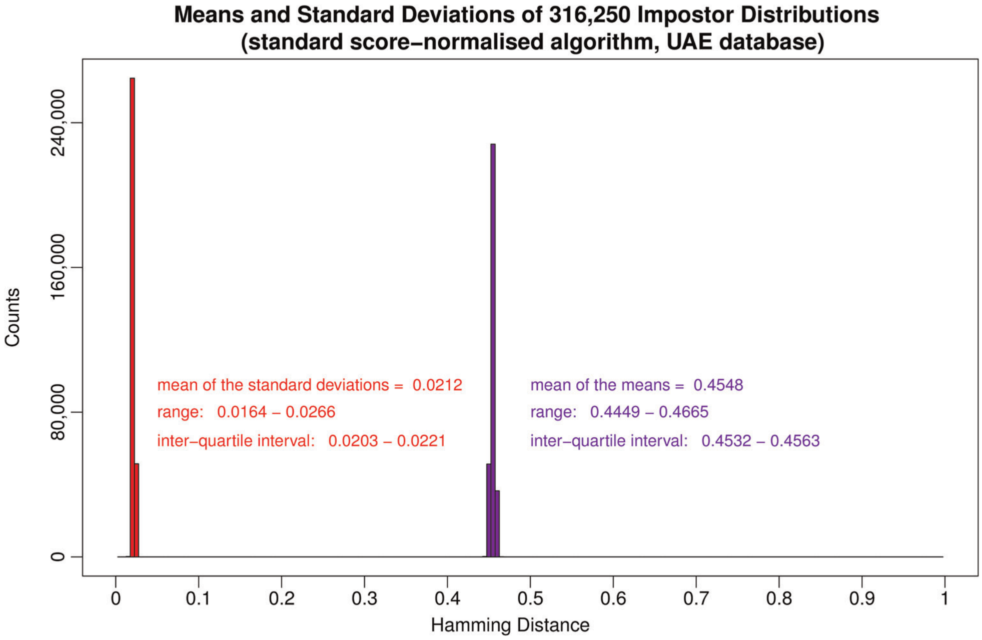

Fig. 2 *Distributions of the means (blue) and of the stnd-devs (red) of 316,250 impostor distributions, each being generated by comparing one eye to 316,250 others. There is very little variation in either statistic among all of these distributions*

distribution, and we do not yet see evidence for a 'Doddington's Zoo' [6] proliferation of types of non-match distributions analogous to Doddington's famous metaphorical wolves and lambs.

4 No wolves, no lambs?

We wish to test this impression of universality to its extremes, by examining the most extreme individual distributions whose scores are described in Figs. 1 and 2. Inevitably, across any distribution of distributions, some will be better and some will be worse. We ordered the 316,250 individual impostor distributions in order of their FMRs at a threshold of HD = 0.35, which is higher than operational deployments (i.e. it leads to worse FMRs), but it allows a meaningful sequencing of the distributions, given the rarity of false matches at lower thresholds even in a database of this size. Clearly, those individual distributions having the lowest mean HD and largest stnd-devs will produce the worst FMRs. The 20 worst distributions with the highest FMR are plotted (colour coded) in Fig. 3, referenced against the canonical (black) distribution of 100 billion impostor comparisons from Fig. 1. In terms of Doddington's Zoo, although the shift from the norm is slight, these may represent 'lambs' (persons, or more precisely eyes, relatively vulnerable to impersonation). Equivalently, given the symmetry of matching, they may represent 'wolves' (relatively successful at impersonation and who prey upon lambs). For the present purposes we cannot distinguish between wolves and lambs because either would appear as a tendency for false matches. In later sections, we will produce detailed inter-quantile analyses of such cases, but meanwhile their extremity should be noted: these are the worst 1/16,000th of the entire distribution of 316,250 distributions.

5 Entropod uniquorns

We have also extracted the 20 best distributions among the 316,250 individual impostor distributions, having the lowest FMRs at threshold HD = 0.35 because their means are highest and their stnd-devs are smallest. These are plotted (colour coded) in Fig. 4, referenced against the canonical (black) distribution of 100 billion impostor comparisons from Fig. 1. Again the shift from the norm is slight, but sufficient to produce zero false matches among their 316,250 comparisons at HD = 0.35 and at HD = 0.36, and nearly zero even at the very liberal HD = 0.37 threshold.

Yager and Dunstone [7] added four new beasts to Doddington's Zoo, all defined in terms of shifts in *both* the impostors and genuine distributions: 'doves' (both distributions move further apart); 'worms' (both distributions move towards each other); 'chameleons' (both distributions move towards greater similarities, whether persons are compared with themselves or with others); and 'phantoms' (both distributions move towards greater *dis*similarities, whether with selves or with others, as phantoms are relatively protean). As the UAE database does not include multiple IrisCodes from individual eyes, we cannot generate distributions of genuine (same eye) scores, and thus we cannot apply to Fig. 4 any of the named beasts from Yager and Dunstone's menagerie with full zoological conformance. In the entertaining tradition begun by Doddington, a new term is needed to describe Fig. 4. As high entropy in a biometric pattern is the origin of its uniqueness, and IrisCodes with higher entropy generate impostor distributions with smaller stnd-devs and means shifted upwards towards 0.5, these two concepts (entropy and uniqueness) should be included in the taxonomy. The coloured distributions of Fig. 4 have migrated toward greater dissimilarity from others on legs of high entropy, and so we propose to name this addition to the menagerie formally as 'entropod uniquorns'.

As a validity check that the division into probe and gallery lists was immaterial, we generated a second impostors distribution for each IrisCode underlying Figs. 3 and 4. This second distribution compared each individual IrisCode not to the gallery but to all the other probes. For each such IrisCode we compared the two distributions and examined the change in mean and variance. These differences were not significant for either group in Figs. 3 and 4. The average paired difference in means was 7×10^{-7} for the entropod uniquorns and -8×10^{-6} for the wolves/lambs (with df = 19, differences of roughly 3×10^{-5} would be required for significance at $\alpha = 0.05$ in a t-test). The average percentage change in variance was −0.13% for the entropod uniquorns and +0.016% for the wolves/lambs (percentage changes of −53% or +73% would be required for significance by χ^2-test).

6 Need to test with score normalisation disabled

Yager and Dunstone [7] wrote that generally, biometric users tend to have individual impostor match score distributions. Fig. 2 seems to

© The Institution of Engineering and Technology 2015

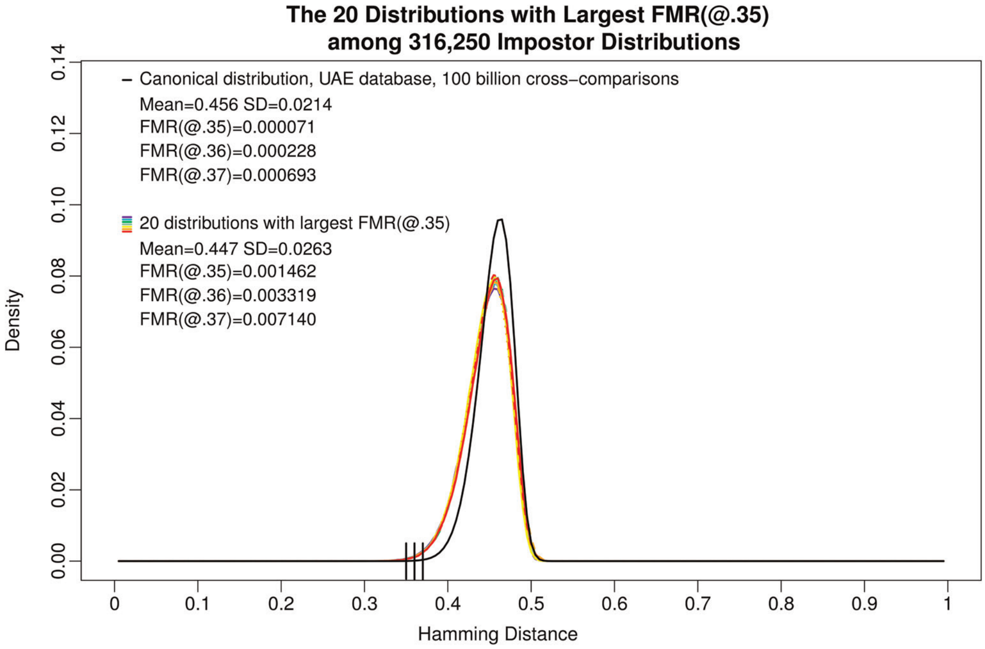

Fig. 3 *Extreme cases of 20 distributions among the 316,250 whose relatively low means and large stnd-devs produce an elevated FMR, analogous to Doddington's wolves and lambs*

contradict this for the vast majority of the 316,250 distributions of iris impostor score distributions for individuals; but at the far extremes of this distribution of distributions, some supporting examples were found (Figs. 3 and 4). Now, we must re-examine this question with a detailed inter-quantile analysis and with any potentially confounding effects of score normalisation eliminated. IrisCodes always disregard (mask out) those bits obtained from the 2D Gabor wavelet projections whose coefficients have amplitudes in the lowest 25% quartile (this longstanding aspect of the IrisCode algorithm has come to be known in some of the more recent literature as 'fragile bit masking'). In addition, any IrisCode bits deemed to have been affected by eyelid or eyelash occlusion, or specular reflections from the cornea or from eyeglasses, are also masked out. When two IrisCodes are compared (in any relative orientation), both sets of masking bits are AND'ed with each other and with the XOR of the data bits, to compute HD for only the mutually unmasked bits. A consequence of these masking operations is that varying numbers of bits are actually compared, and this variation if ignored would affect the HD score distributions. If large numbers of bits from unrelated IrisCodes are compared, we would expect their HD score to approach 0.5, for the same reason that tossing a fair coin many times in a run is unlikely to produce large deviations from a 50/50 outcome. Likewise, if only relatively few bits survive these mutual masking operations, a spuriously low HD score could arise just by chance, for the same reason that tossing a fair coin few times can readily yield all one outcome.

Therefore, the standard publicly deployed algorithm for iris recognition utilises score normalisation to compensate for variation

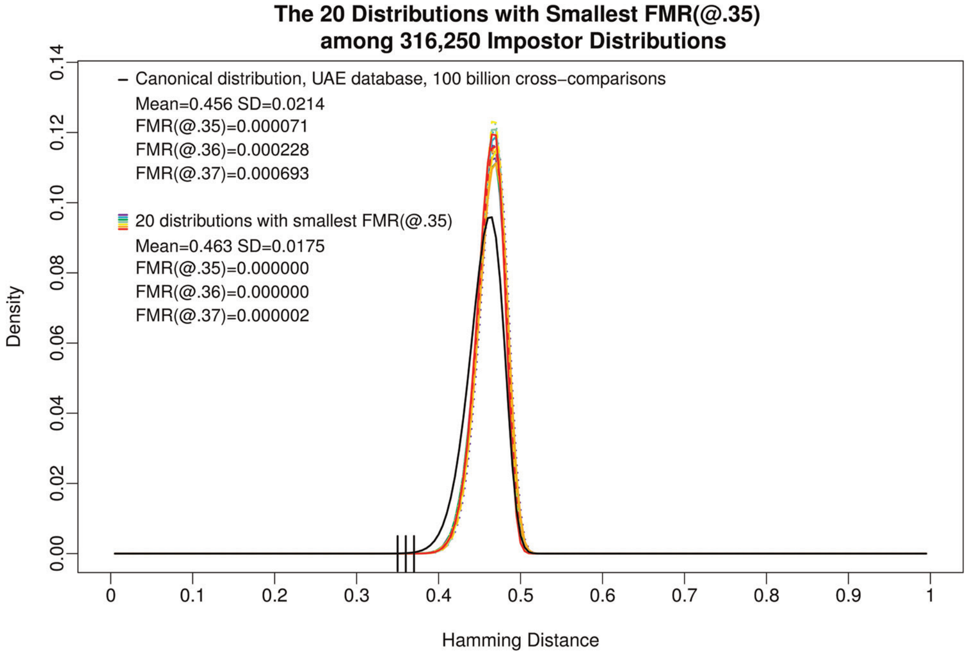

Fig. 4 *Extreme cases of 20 distributions among the 316,250 whose relatively high means and small stnd-devs produce an especially benign FMR, indeed zero FMR at the thresholds 0.35 and 0.36 HD. They join Doddington's Zoo as novel creatures: 'entropod uniquorns'*

© The Institution of Engineering and Technology 2015

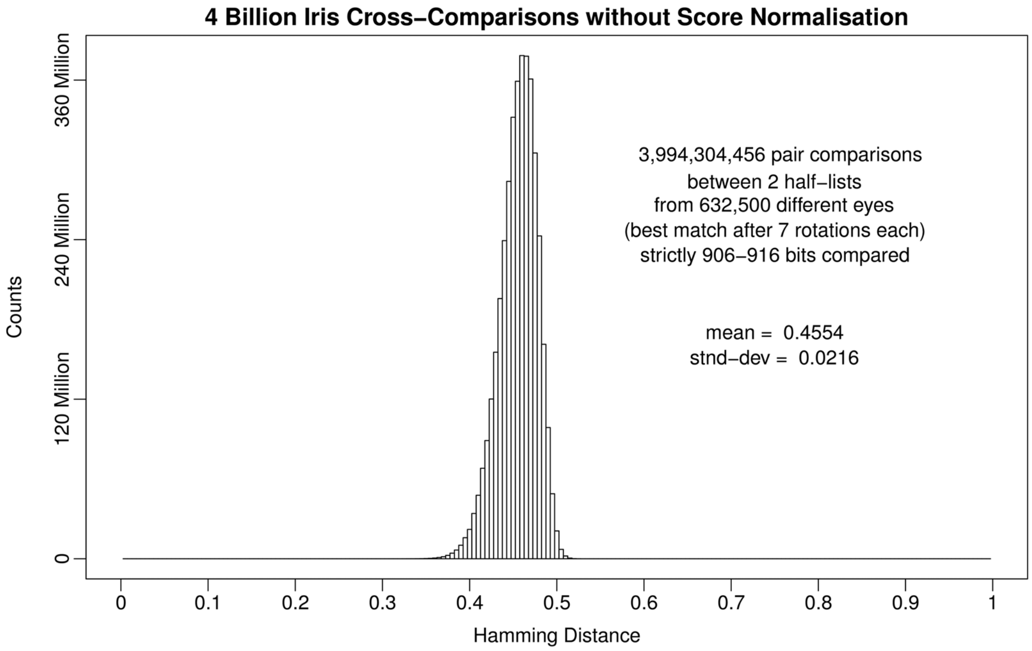

Fig. 5 *Among 100 billion pairings between probe and gallery IrisCodes in the UAE database, 4 billion fell into a narrow window of $906 \leq n \leq 916$ bits mutually unmasked for comparison. This resulting impostor distribution of HD scores is almost indistinguishable from the full set in Fig 1, but by seeking individual differences within only this isomerous subset of pairings we can rule out any potential contaminating effects of score normalisation*

in the number of bits actually compared between any two IrisCodes. Raw fractional HD scores between unrelated IrisCodes based on n bit comparisons tend to be distributed with their stnd-dev from 0.5 inversely proportional to $\sqrt{n}$ (before selecting the lowest after several rotated comparisons), and typically about $n = 911$ bits are mutually unmasked for comparison. Therefore, deviations from HD = 0.5 when n bits were compared are rescaled by $\sqrt{n/911}$ whether $n > 911$ or $n < 911$ to generate a normalised score, thus enabling commensurability of match scores regardless of n. In effect, if $n < 911$ bits were compared then a higher match quality is required, but if $n > 911$ then the match standard becomes more forgiving, while still allowing all such matches to be comparable with each other. For the present investigation, however, this normalising mechanism might either conceal individual differences in impostor score distributions, or indeed it might even be the source of those extreme cases revealed in Figs. 3 and 4. Therefore, we must disable the score normalisation, and examine the question again for just those IrisCode pairings in which comparable numbers of bits n were mutually unmasked and compared.

Fortunately, the vast number of pairings possible between unrelated IrisCodes in the UAE database is so large that we can restrict consideration to only those comparisons in which a narrow window of 911 ± 5 bits were mutually unmasked, and for these cases we can disable score normalisation. To generate all results

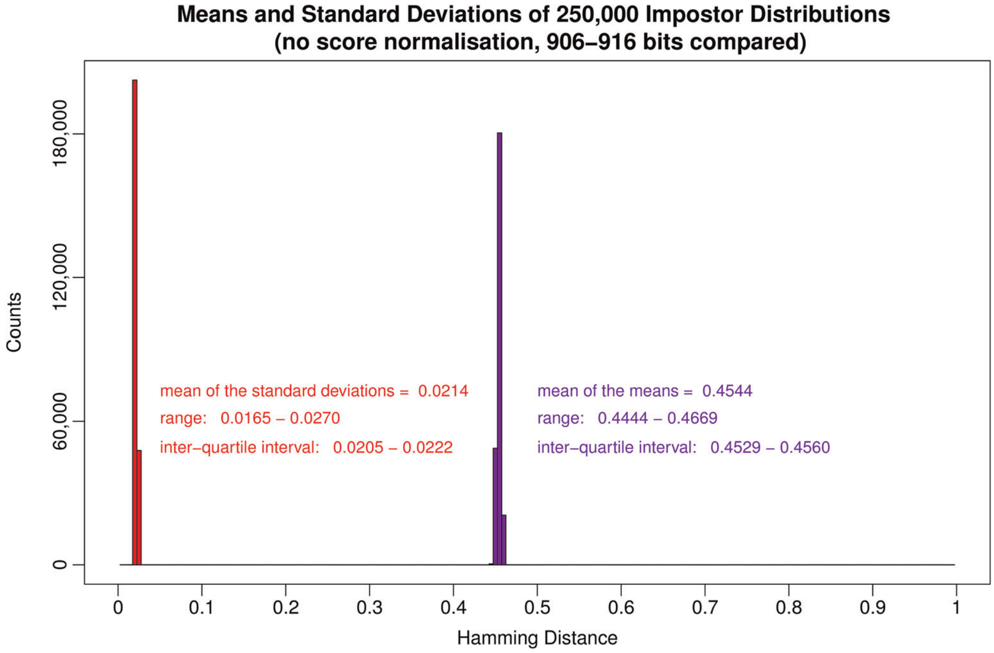

Fig. 6 *Distributions of means (blue) and of stnd-devs (red) of HD scores between unmated IrisCodes selected both for the narrow window of 911 ± 5 bits mutually unmasked for comparison, and for having at least 5000 such encounters with other eyes in the gallery to generate meaningful distributions of scores*

© The Institution of Engineering and Technology 2015

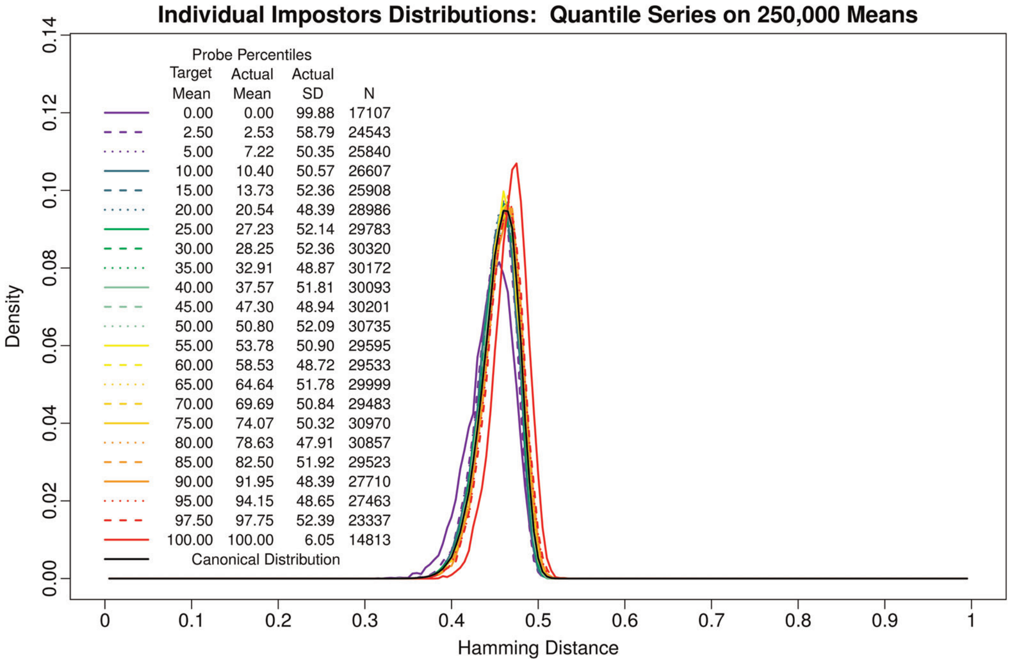

Fig. 7 *Inter-quantile sequence of impostors distributions ordered by mean HD scores. Almost no separation is apparent between the 2.5th percentile and the 97.5th percentile distributions*

presented in the rest of this paper, we again compared one half of the UAE database against the other, but we ignored any comparisons for which the requirement that $906 \le n \le 916$ bits be mutually unmasked was not satisfied. Among the 100 billion IrisCode comparisons generated and described in Figs. 1 and 2, about 4% (3.994 billion) passed through this window. That distribution of impostor score comparisons (with score normalisation disabled), which we may term *isomerous* comparisons because of the nearly equal numbers of bits compared when generating all the scores, is shown in Fig. 5. It appears almost indistinguishable from the full distribution incorporating score normalisation that was plotted in Fig. 1. However, since we now wish to study the *distribution of distributions* of such unnormalised impostor scores each generated by an individual probe, to search (as before) for individual differences, we must further select only such distributions in which a sufficient number of encounters occurred with other eyes in the gallery (>5000) to allow meaningful distributions to be generated and analysed. Applying this second filter to the 316,250 probe IrisCodes, 250,279 of them still survived. Those quarter-million individual distributions of isomerous scores had the distribution of means shown in blue in Fig. 6.

7 Homoscedasticity

The distribution of mean HD scores in Fig. 6 is almost indistinguishable from the distribution of mean HD scores in Fig. 2 for the 316,250 distributions computed *with* score normalisation. Both distributions of means are extremely narrow, supporting the thesis of a nearly universal form for the IrisCode impostors distribution. Similarly, the distribution of stnd-devs for the 250,000 isomerous distributions shown in red in Fig. 6 is very narrow, and almost indistinguishable from the score normalised stnd-devs in Fig. 2. Magnified versions of the histograms shown in Figs. 2 and 6 are available online at [8], along with QQplots against normal and χ^2 distributions. Means are generally well fit by a normal distribution and variances by a χ^2 distribution. In the absence of the images underlying the distributions, which were not stored upon enrolment in this database, it may remain unclear whether the observations at the tails of these distributions are statistical anomalies. Future work on other databases which *do* include the original images will seek to determine whether image properties or iris properties produce the impostor distributions having the extremes of means and of stnd-devs.

Both in Fig. 2 and in Fig. 6, we see evidence of approximate *homoscedasticity:* a set of distributions, or a sequence of random variables, having the same variance is homoscedastic. If IrisCodes produced impostor match score distributions whose stnd-devs differed significantly among individuals, as Yager and Dunstone [7] suggested was generally true across biometric modalities, then Figs. 2 and 6 would not be such narrow spikes, and the score distributions would be *heteroscedastic*. To examine the universality, and particularly the homoscedasticity, of the impostors distribution in greater detail, we must now apply inter-quantile analyses to these 250,000 isomerous distributions. As discussed earlier, understanding the universality of the impostors distribution is important for optimal search strategies and for interpreting match scores.

8 Detailed inter-quantile analyses without score normalisation

We ordered the 250,000 isomerous distributions of impostor scores first in order of their mean HD scores. This ordering allowed us not only to extract the ones having minimum and maximum means, but also to examine the separation between distributions at a series of quantiles between 2.5% and 97.5%. For plotting purposes we selected distributions within 2.5% of the target mean quantile and as close as possible to the median stnd-dev. For most mean quantile bands, we were able to select an individual distribution within 2.5% of the stnd-dev median, although this was not possible at the extremes of the range (mean quantiles of 0%, 2.5%, and 100%). Twenty-three such distributions spanning the full range of means quantiles are plotted in Fig. 7, as well as the canonical distribution from Fig. 5 (black). Tests confirmed that, once again, the form of these distributions did not depend on the partitioning of the database into probe and gallery sets. Paired comparison significance tests demonstrated this for both the kinds of quantile series presented in Figs. 7–10.

We see in Fig. 7 that there is almost no separation among any of the distributions between the 2.5th percentile and the 97.5th percentile, and that they are nearly indistinguishable in closely hugging the canonical form; but at the very extremes a gap does

© The Institution of Engineering and Technology 2015

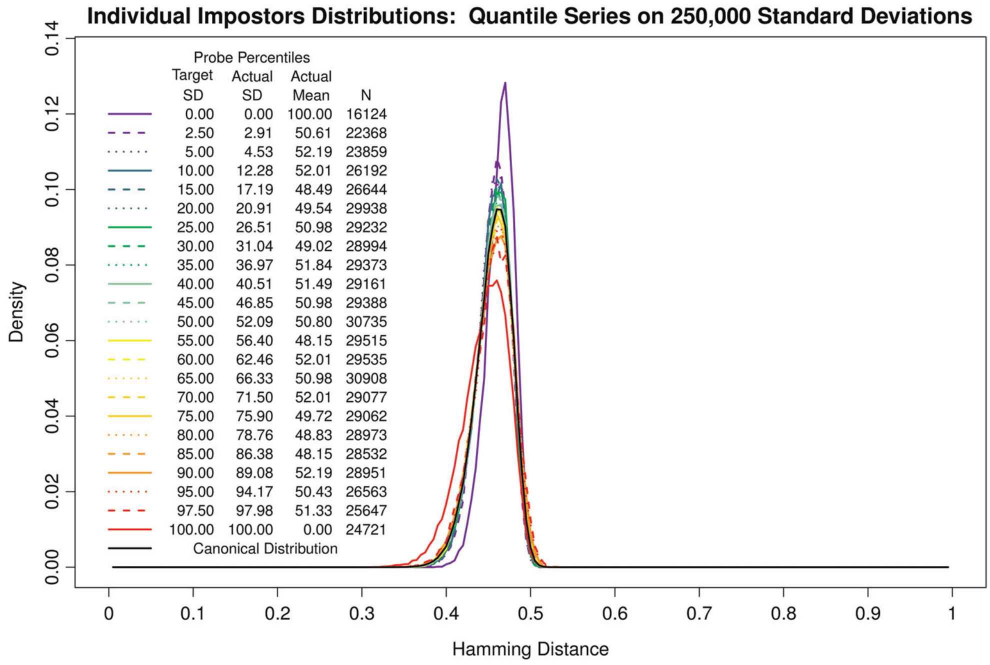

Fig. 8 *Inter-quantile sequence of impostors distributions ordered by their stnd-devs. Almost no separation is apparent between the 2.5th percentile and the 97.5th percentile distributions*

appear before the 0th percentile (blue, minimum) and before the 100th percentile (red, maximum) outlier distributions. Data tabulated within Fig. 7 gives further properties for each of the plotted distributions: its actual mean percentile (for comparison to the target quantile); its actual stnd-dev percentile (again for comparison to the target 50% quantile); and the number of HD scores within each isomerous distribution, which we maximised in selecting from each quantile band.

We also ordered the 250,000 isomerous distributions of impostor scores by stnd-dev, and we selected for plotting those distributions falling within specified quantile bands on stnd-dev, with mean scores as close as possible to their median (ideally ±2.5%) and otherwise maximising the number of scores available. Twenty-three distributions representing various percentile quantiles in this ordering are plotted in Fig. 8, as well as the canonical distribution from Fig. 5 (black). Again in Fig. 8 we see almost no separation among the distributions nor from the canonical form except for the extreme 0th percentile (blue) and the 100th percentile (red) distributions. It is noteworthy that these two outliers have the property that the distribution with largest stnd-dev (solid red) has the lowest mean, and the distribution with the smallest stnd-dev (solid blue) has the highest mean. A similar relationship was observed in Fig. 7, and we shall study this property later in more detail.

We now probe those outermost quantile regions beyond the 97.5th percentile and below the 2.5th percentile, for both the means and stnd-devs of the 250,000 isomerous impostor distributions. Highly

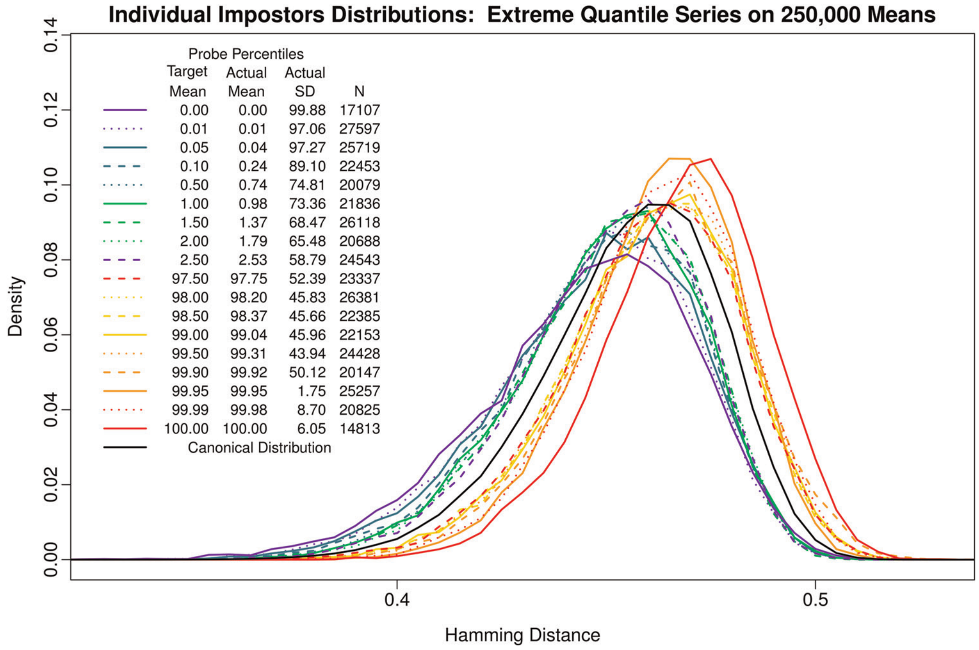

Fig. 9 *Greatly magnified inter-quantile sequence of impostors distributions ordered by mean HD scores, for the distributions within the outermost 2.5th percentile quantiles*

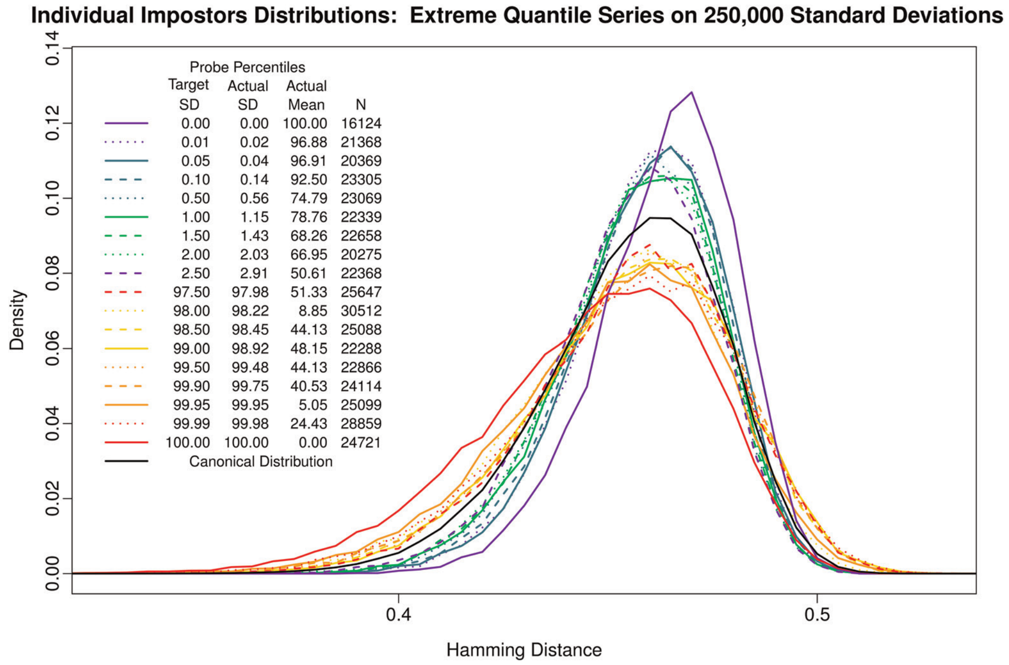

Fig. 10 *Greatly magnified inter-quantile sequence of impostors distributions ordered by their stnd-devs, for the distributions within the outermost 2.5th percentile quantiles*

magnified plots of the distributions representing those extreme quantiles are presented in Figs. 9 and 10, respectively. The rarity of such departures from the canonical impostors distribution is very striking: it is only at these extremes that departures from the canonical impostors distribution are visible. It is also noteworthy that beyond the outer 2.5% quantiles at both extremes, both for means and for stnd-devs, it was generally impossible to find a distribution in which the other parameter was near its median. Rather, an inverse relationship developed between the quantiles of the two parameters. This effect is evident numerically in the actual mean and stnd-dev quantiles listed in these figures, but we can document the relationship more fully by explicitly analysing the dependence between these two parameters.

For this analysis, each of the 250,000 impostor distributions were assigned to one cell of a (100 × 100) grid defined jointly by the 1% quantiles determined separately for the means and the stnd-devs. By construction, the sum of the number of distributions assigned to cells in each row, and in each column, of the grid is constant (at 1% or 2500). Further, if the means and stnd-devs are independent, a fixed number of distributions (0.01% of the 250,000) would be expected to be assigned to each cell. If one were to depict such a grid with cell counts coded by colour, and if means and stnd-devs were independent, the expected image would be a uniform field of only one colour. However, to the extent that the two parameters were dependent, the image produced would be more structured and multi-coloured. Such a depiction of the present data is shown in

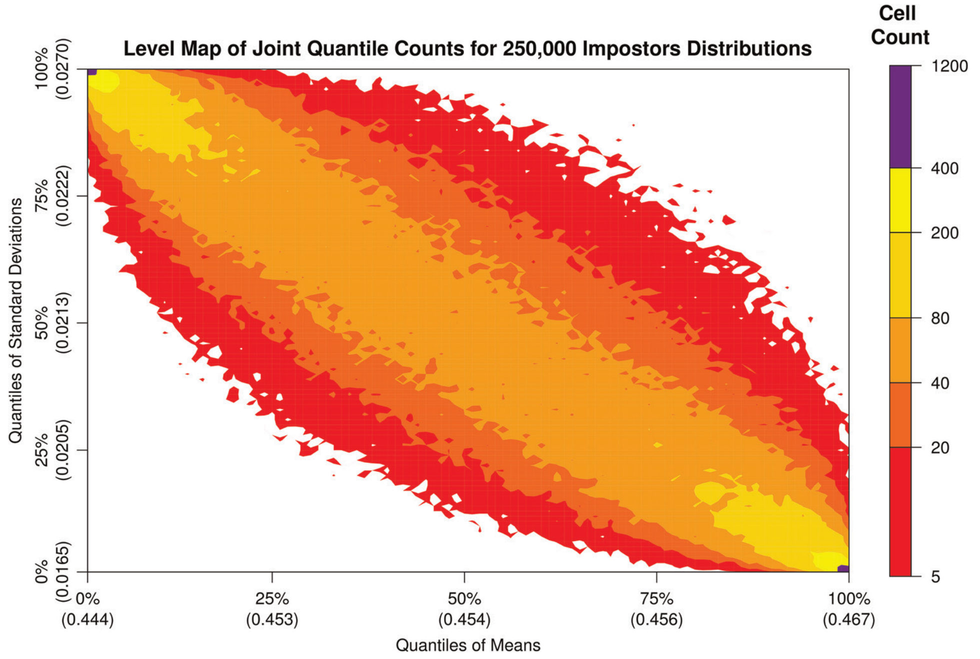

Fig. 11 *Colour-coded level map of joint quantile counts of means and stnd-devs of all 250,000 distributions of impostors scores, revealing a hidden inverse dependency*

© The Institution of Engineering and Technology 2015

Fig. 11. It reveals a striking inverse relationship between the means and stnd-devs of these distributions, as does a standard calculation of their correlation ($r = -0.86$, $p < 2.2 \times 10^{-16}$).

Impostors distributions with higher means tend to have smaller stnd-devs (both of which reduce FMR), and those distributions with lower means tend to have larger stnd-devs (both of which increase FMR). We also see at the poles in Fig. 11 two concentrations of distributions that are maximally and inversely extreme on each dimension. These two polar concentrations are prototypical entropod uniquorns, at one extreme, while at the other they are analogous to Doddington's wolves or lambs. Thus, although our primary observation is that impostors distributions associated with different individual probes tend to be much the same, with means and stnd-devs confined to a narrow range as shown in Fig. 6, we note in Fig. 11 a striking inverse relationship within that narrow range. We speculate that this reflects variation in the entropy of different iris patterns, because as noted in Section 5, such variation would produce these observed joint effects. It is remarkable that this occurs even when the number of bits compared for each HD score is confined to such a narrow window of 911 ± 5 mutually unmasked bits, so variation in data length cannot explain this apparent variation in entropy.

To characterise further the differences between the canonical and the individual impostors distributions examined, we performed one-sample t- and χ^2-tests on the differences between the mean and variance of each distribution and the appropriate population means and variances given in Figs. 2 and 6 for score-normalised and non-normalised cases. Generally, we can reject the null hypothesis (of no difference) when the mean or stnd-dev departed from its median by a ventile or a decile. For the experimental variable, this occurred when we intentionally selected distributions to fall within a non-central quantile band. For any non-experimental variable, it occurred when we were not able to find a distribution that was both extreme on the experimental variable and central on the non-experimental variable (and we therefore sacrificed centrality on the non-experimental variable). Very small differences sufficed for significance. To achieve significance at $\alpha = 0.05$, means needed to change by only about 0.00006 and 0.0002 for score-normalised (with larger N) and non-normalised distributions, respectively. Variances needed to change by only 0.4% and 1.6%, respectively.

These are essentially tests of whether the selected distributions could be random samples from the two populations of impostors distributions described. The significance of these tests is therefore not a surprise, since we intentionally selected distributions to represent different quantile positions or extremes from the population, and it may function more as a validity check on our methodology. These findings of significance also arise as a consequence of the fact that with very large numbers of observations, even very small differences can be statistically significant, although not necessarily meaningful. In view of this, we focus instead on the meaningfulness of the differences observed.

9 Implications for search strategies: 1-to-1, 1-to-first, 1-to-many

We turn now to the significance of these observations for matching strategies. As mentioned in the Introduction, the interpretation of a given match score is greatly simplified if it can be assumed, at least to a very good approximation, that there exists a universal and known distribution for impostors scores. We shall first consider several operational modes and decision strategies when the assumption is made that there is an essentially universal form of impostors distribution.

9.1 1-to-1 verification

The simplest case is 1-to-1 verification, although iris recognition is rarely deployed in such an undemanding mode. Assuming there is a known universal probability distribution $f_1(x)$ for the dissimilarity scores x that impostors generate (in the present case x is the smallest fractional HD obtained after comparing two IrisCodes in the requisite number of rotations for tilt, so $f_1(x)$ is as exemplified by the data plotted in Figs. 1 and 5), then the probability $F_1(x)$ of making a false match in single verification trials when the decision rule is to accept any score of x or less, is the cumulative under the probability density distribution up to x:

$$F_1(x) = \int_0^x f_1(x)dx \tag{1}$$

or equivalently

$$f_1(x) = \frac{d}{dx}F_1(x)\,. \tag{2}$$

9.2 1-to-first, or 1-to-many, identification

Iris recognition is almost always deployed in identification mode, meaning that an observed score is accepted as a match only if it passes some rather demanding threshold x after an exhaustive search through a database of N persons; or alternatively a match is declared the *first* time the score threshold test is passed and then the search is stopped. Both in the 1-to-first and in the exhaustive search strategy, let N be the total number of impostor scores computed. Then the probability of *not* making a false match on *any* of those N opportunities is $[1 - F_1(x)]^N$, where $F_1(x)$ is as defined in (1), and so the net false match probability $F_N(x)$ is

$$F_N(x) = 1 - [1 - F_1(x)]^N \tag{3}$$

which can be approximated as $F_N(x) \simeq NF_1(x)$ where $F_1(x)$ is very small: $F_1(x) \ll 1$ provided that $[NF_1(x)]^2$ also remains small (binomial theorem). In other words, 1-to-first or 1-to-many identification on N is roughly N times more prone to false matches than mere 1-to-1 verification which poses only a single opportunity of error. Fortunately, the IrisCode has been confirmed by NIST [9] in tests involving 1.2 trillion iris comparisons to have such minuscule false match probability that it easily tolerates this N-fold increase in the required resilience when searching a database of N persons. For example, at an HD score threshold of $x = 0.28$, allowing 28% of bits to disagree, NIST reported [9] a false match probability of $F_1(x) = 1$ in 40 billion.

For completeness, the probability density distribution $f_N(x)$ associated with the cumulative false match probability $F_N(x)$ after a 1-to-first or a 1-to-many search involving N impostors is

$$f_N(x) = \frac{d}{dx}F_N(x) = N[1 - F_1(x)]^{N-1} f_1(x). \tag{4}$$

Out on the left tail of the density where $f_1(x)$ and therefore $F_1(x)$ are very small, we note the approximation that $f_N(x) \simeq Nf_1(x)$ as a reflection of the N-fold greater resilience needed when making N impostor comparisons in an identification search.

9.3 Biometric 'birthday problem'

We return now to the biometric 'birthday problem' with which this paper opened. A biometric technology operating at threshold x with a FMR of $F_1(x)$ becomes more likely than not to produce at least one biometric collision among a group of N persons when

$$[1 - F_1(x)]^{N(N-1)/2} < 0.5 \tag{5}$$

and it is easily shown that if $F_1(x)$ is small, this condition is satisfied once $N > \sqrt{1.386/F_1(x)}$.

© The Institution of Engineering and Technology 2015

9.4 *If a universal form of impostors distribution cannot be assumed*

The above analyses all rely on the existence of a universal impostors distribution $f_1(x)$, and more particularly on its cumulative $F_1(x)$. The analyses that now follow support the validity of this assumption for all individual distributions except those at the extremes of the quantile spectra. In those outer regions, to avoid an unwanted impact on FMR, it may be advisable to incorporate a subject-specific $f_1(x)$ and $F_1(x)$ above for the various operational search and decision strategies.

10 Conclusions: the universal and the particular

We have seen that individual distributions remain quite faithful to the canonical distribution and only depart from it when they are extreme in mean and stnd-dev. This comparison can be refined in an informative way by determining the extent to which the canonical impostors distribution might over- or underestimate individual FMRs over the range of quantiles considered earlier.

For the canonical distribution and for each of the 250,000 individual distributions, we calculated FMRs at a threshold of HD = 0.37 (chosen so high to ensure that false matches do actually occur despite having only about 10^4 scores in each individual distribution). We then calculated the ratio of each individual FMR to the canonical FMR (which was 0.000759 at HD = 0.37).

Fig. 12 summarises this elevation or reduction factor for each of the quantile bands used previously in Figs. 7–10. The coloured bars indicate the median FMR elevation or reduction factor (red or green) for all of the individual distributions within that band, and the whiskers show the inter-quartile interval (first and third quartiles). The plotted triangles also show the median for that subset of distributions within each quantile band that fall within 2.5% of the median on the other quantile variable.

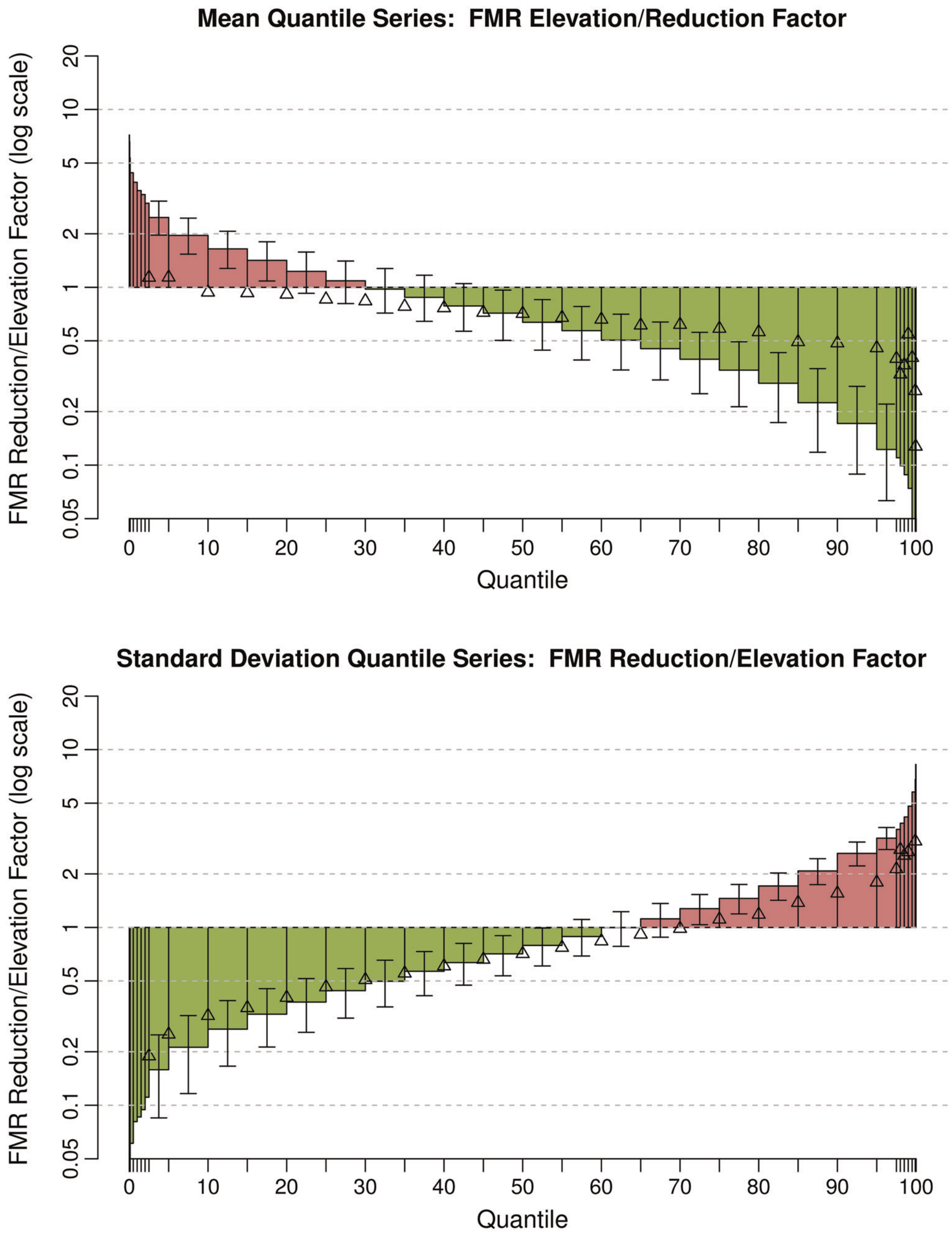

Fig. 12 *Elevations and reductions of FMR at HD = 0.37 relative to the canonical FMR for subpopulations of individual distributions along the quantile spectra. Important elevations of FMR occur only when both mean and stnd-dev are highly extreme. Bars and whiskers indicate subpopulation medians and inter-quartile intervals. Triangles indicate medians for subpopulations further restricted to remain central on the other quantile variable. Ordinate is logarithmic*

© The Institution of Engineering and Technology 2015

From this figure, we can see that the canonical FMR at HD = 0.37 would substantially underestimate individual FMR only very rarely. Furthermore, we can use Fig. 12 to calculate the expected frequency of underestimates by different magnitudes. We can see: that an underestimate by more than a factor of 10 would almost never occur (in fact it occurs for 0.0036% of the 250,000 individual distributions); that it would occur by more than a factor of 5 for only the most extreme quantiles of the mean and stnd-dev series (actual frequency 0.57%); by more than a factor of 3 in only the most extreme 2.5–5% (actual frequency 4%); and by more than a factor of 2 for the most extreme 10–15% (actual frequency 12%). We can also see that the canonical distribution would underestimate FMR, to any extent whatsoever, for only about 35% of the individual distributions (actual frequency 36%).

Thus, from Fig. 12 we can conclude that for the vast majority of individual distributions, the canonical distribution would not underestimate FMR, and that where it did, the magnitude of the underestimate would be operationally insignificant for all but a very rare population of possible ‘wolves/lambs’ (the most extreme red bars plotted in Fig. 12). For persons fortunate enough to be ‘entropod uniquorns’ (the most extreme green bars plotted in Fig. 12), their reduced actual FMR is simply a bonus and of no concern. We conclude that apart from those rare exceptions, whose rarity we have calibrated here, the IrisCode generates a remarkably invariant impostors distribution. The large entropy which lies at the heart of this biometric technology [5] confers on it the key advantage of having a safe and relatively constant dissimilarity distance when different persons are compared, and thereby the absence of doppelgängers.

11 Acknowledgment

Statistical analyses and generation of figures were performed using the ‘R’ package: https://cran.r-project.org/.

12 References

1 http://www.CL.cam.ac.uk/users/jgd1000/Doppelganger-photos.pdf

2 https://portal.uidai.gov.in/uidwebportal/dashboard.do, Indian Government dashboard showing enrollment progress of the Unique IDentification Authority of India, updated weekly.

3 Daugman, J.G.: ‘Probing the uniqueness and randomness of IrisCodes: results from 200 billion iris pair comparisons’, *Proc. IEEE*, 2006, **94**, (11), pp. 1927–1935

4 Daugman, J.G.: ‘High confidence visual recognition of persons by a test of statistical independence’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993, **15**, (11), pp. 1148–1161

5 Daugman, J.G., Downing, C.J.: ‘Epigenetic randomness, complexity, and singularity of human iris patterns’, *Proc. R. Soc. Lond. B, Biol. Sci.*, 2001, **268**, pp. 1737–1740

6 Doddington, G., Liggett, W., Martin, A., *et al.*: ‘Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation’. Proc. of Int. Conf. on Spoken Language Processing, 1998

7 Yager, N., Dunstone, T.: ‘The biometric menagerie’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (2), pp. 220–230

8 http://www.CL.cam.ac.uk/users/jgd1000/SupplementaryGraphsDoppel.pdf

9 Grother, P., Quinn, G.W., Matey, J.R., *et al.*: ‘IREX-III: performance of iris identification algorithms’. NIST Interagency Report 7836, NIST, Gaithersburg, MD, April 6, 2012

© The Institution of Engineering and Technology 2015