Model of the Behavior of the IBIS Correlation Scores in a Large Database of Cartridge Cases

By: Dr. Alain Beauchamp and Dr. Danny Roberge, Forensic Technology, Montréal, Québec, Canada [1]

Keywords: IBIS, large database

Abstract

Nearly 500 pairs of cartridge cases, with each pair fired from the same firearm, have been acquired with IBIS™, for each of the following four calibers: 9mm, 32 Auto, 45 Auto and 22. The acquired marks are the breech face mark, the ejector mark, and the firing pin mark for the first three calibers and the rimfire firing pin mark for the last caliber. Furthermore, the ejector mark has been acquired for 78 pairs of 9mm cartridge case. Each acquired cartridge case was correlated against the other acquired images in the database, including the corresponding ("sister") cartridge case from the same pair. The resulting score lists have been used to create a statistical model that simulates the performance of the IBIS system when using a large database of cartridge cases.

The score distribution that is associated with a given reference and the rest of the database (excluding the reference's sister) has a uniform shape that is independent of the reference itself, if the scores are properly normalized. In the model, a smooth statistical distribution is generated by combining the normalized non-sister scores, and by extrapolating the available data in the right wing by an analytical function. This last step is crucial since the probability of finding matching pairs in a large database is sensitive to the shape of the distribution in that region.

In this work, a given reference cartridge case R and its sister are defined as a "found matching pair" (a hit) if their correlation score is among the l0 largest scores between R and all cartridge cases with the same caliber in the database. From this definition, the performance of the algorithm is given by the proportion of found matching pairs. A performance curve, that predicts the performance of the IBIS system as a function of the database size *for the acquired sister pairs* that are used to train the model, is computed. The expected performance decreases from 80% to 30–45% when the database size increases from 1000 to one million exhibits. The model also confirms that acquiring all available marks improves the combined performance. The breech face, firing pin and ejector marks thus provide complementary information. The model and tests results provided in this paper are based on non filtered databases of raw images. The actual number of test samples in a database could be significantly lower in real life as the databases themselves could be segmented into smaller sample sets. The segmentation of the database based on geography or type determination characteristics of the firearm could reduce the correlation database significantly.

Error analysis was also performed for the 9mm, 32 Auto and 45 Auto calibers, with the breech face and firing pin marks. In every case, the predicted correlation performance computed with the model, when trained with a small database (about 1000 exhibits), agreed with the actual performance in a larger database (4000 to 56,000 exhibits), within (or nearly within) the error bars.

1. Introduction

A critical issue about the feasibility of a national ballistic imaging system is its performance for a large database with millions of exhibits. Since no such database actually exists, tests have been performed at Forensic Technology (FT) in 2004 and 2005, on smaller databases of cartridge cases for selected calibers (9mm, 32 Auto, 45 Auto, and 22). In these tests, nearly 500 pairs of cartridge cases (each pair fired from the same firearm) were acquired with the IBIS system, and introduced into a larger database (4000 to 56,000 exhibits of the same caliber).

In FT's tests, the performance of the algorithm is determined as follows. At first, each member of the pairs acquired at FT (from now on, called sister pairs) is set as a "reference" and is compared with the other cartridge cases in a database that contains 500 pairs. This process is loosely referred to as "correlation" within the IBIS users' community. A list of scores[2] is then associated with each reference cartridge case; the score of the reference's sister is within the list. For each reference,

[1] Senior scientists, alain.beauchamp@contactft.com, danny.roberge@contactft.com

[2] A high score indicates a high similarity between two cartridge cases.

the list of scores is sorted in decreasing numerical order, and the rank of the sister's score is computed. At this point, the reference's sister is defined as a found match if its rank is less than or equal to some integer M. In practice, IBIS users often consider exhibits with the 10 largest scores as good match candidates in their everyday work and, similarly, M = 10 has been adopted here. When this operation is completed for all references, the performance of the algorithms is given by the proportion of sisters that are found as matches according to the previous definition. The performance that was measured in FT's studies was scaled from 50 to 90%. While very valuable, these studies, as such, cannot predict performance for large database sizes that could characterize a national ballistic imaging system.

The goal of this paper is to present a statistical model that has been developed at FT to predict the performance of the IBIS correlation algorithm as a function of the database size. More precisely, we want, from a set of known sister and non-sister pair scores, to predict the proportion of *those* sister pairs that would be found as a match as a function of the database size. The conclusion inferred from the model should be applied only to the sister pairs that are used as input. It is not our intent to predict the performance of the model for some hypothetical sister pairs with no available scores. Different numerical results could be obtained, even with a same given caliber, if the sister pairs that are input into the model are purposely selected with a low or high level of similarity as determined by an "eye" inspection. It is also assumed that the non-sister score distribution in the *simulated* database is the same as in the *experimental* database.

Sections 2, 3 and 4 cover the different phases in the development of the model, while section 5 presents numerical results. Section 2 introduces the method that is used to compute the probability that a given sister is found as a match, as a function of the database size. It is shown that this probability decreases as the size of the database increases. The formal definition of the performance curve is also given. Section 3 describes the strategy that is used to derive a "universal" normalized non-sister score distribution, called as such since it does not depend on the reference cartridge case. A good knowledge of that distribution is required to compute the probabilities in the model. Error analysis is considered in Section 4. Section 5 then validates the model using experimental results and gives the expected performance of the IBIS system for four calibers, as computed from the available data that was used in the previous tests performed at FT. Conclusions are presented in Section 6.

2. Basics

This section describes the first phase in the development of the model. Mathematical expressions are derived to compute the probability that a reference A and its sister A' are found as a matching pair as a function of the database size. We then give a simple analytical expression for the critical database size at which that probability degrades significantly from unity. The performance curve that will be used to quantify the performance of IBIS for a large database is also defined. In this section, we assume that the non-sister score distribution is perfectly known. Section 3 will describe the method that is used to determine the distribution from the available data.

2.1. Probability that a Sister is Found as a Match as a Function of Database Size

Let A-A' be a sister pair of a given caliber and $S_{AA'}$ be the corresponding score that measures their relative similarity for a given mark type T (e.g., breech face, firing pin or ejector mark of cartridge cases). Furthermore, let f(x; A, T) be the probability distribution of the scores between A and a pool of non-sisters of the same caliber. We assume that the pool is very large so that millions of cartridge cases could be selected randomly from it.

The probability that a non-sister cartridge case that is selected randomly from the pool obtains a score x which is greater than $S_{AA'}$ when correlated with A is given by

$$Prob(x > S_{AA'}; A,T) = \int_{S_{AA'}}^{\infty} f(x'; A,T)dx' = 1 - C(S_{AA'}; A,T),$$

where $C(S_{AA'}; A, T)$ is the cumulative distribution,

$$C(S_{AA'}; A,T) = \int_{-\infty}^{S_{AA'}} f(x'; A,T)dx'$$

which depends, in full generality, on the reference cartridge case A and mark type T.

Now, consider the following random experiment: a database D is built by randomly selecting N exhibits from the pool of A's non-sisters. The A's sister is then introduced into that database, and a correlation is performed between A and each element of D (including A'). The probability that the sister score $S_{AA'}$ is at the M+1th position in the sorted list of scores is found by computing the probability that M non-sisters obtain a score that is greater than $S_{AA'}$. This step involves the binomial distribution, which is described next.

Let the output of a random experiment be classified in only two mutually exclusive ways, success or failure, with probability p and 1-p, respectively, and let M be the random variable that measures the number of successes from N (M) independent trials. It is well known that M follows the binomial distribution

$$B(M) = \frac{N!}{M!(N-M)!}\, p^{M}\ (1-p)^{N-M}$$

with average Np and variance Np(1-p).

Returning to the original experiment, if obtaining a non-sister score that is greater than the sister score $S_{AA'}$ is defined as success, the probability of success is p = 1-C($S_{AA'}$). From the binomial distribution, the probability that M non-sisters obtain a score that is greater than S in N trials or, equivalently, that the rank of A' is M+1 in the sorted score list, is found to be

$$Prob(SisterRank = M+1; A, T, N) = \frac{N!}{M!(N-M)!}[1-C(S_{AA'}; A, T)]^{M}\ C(S_{AA'}; A, T)^{N-M}\ .$$

In particular, the probability that A' will be in first position (M = 0) and within the first 10 positions (M ≤ 9) is

$$Prob(SisterRank = 1; A, T, N) = C(S_{AA'}; A, T)^{N}$$

and

$$Prob(SisterRank \le 10; A, T, N) = \sum_{M=0}^{9} \frac{N!}{M!(N-M)!}[1-C(S_{AA'}; A, T)]^{M}\, C(S_{AA'}; A, T)^{N-M}$$

respectively.

2.2. Behavior of the Sister Rank for an Academic Case

The previous expression for the probabilities predicts an unavoidable degradation in the rank of the sister score as the database size N increases. Figure 1 (left) shows an analytical distribution of non-sister scores with some exhibit A, here a gaussian distribution normalized to null mean and unit variance for convenience, along with a dotted vertical line, passing through a (hypothetical) score $S_{AA'}$ = 4 that is associated with A's sister. Note that this score value is clearly in the far right wing of the distribution.

The right part of Figure 1 shows the cumulative distribution of the non-sister scores that is "elevated" to the 1st, 1000th and 1,000,000th power (case 1, 2 and 3, respectively). These three functions are the probability distribution corresponding to obtaining a sister score $S_{AA'}$ in first position when the sister is introduced in a database of 1, 1000 and 1,000,000 non-sisters, respectively. It is seen that score $S_{AA'}$ has a 100%, 95% and nearly 5% probability of being in the first position in the three different databases. Despite a high sister score value, the sister rank is expected to degrade at sufficiently large database sizes.

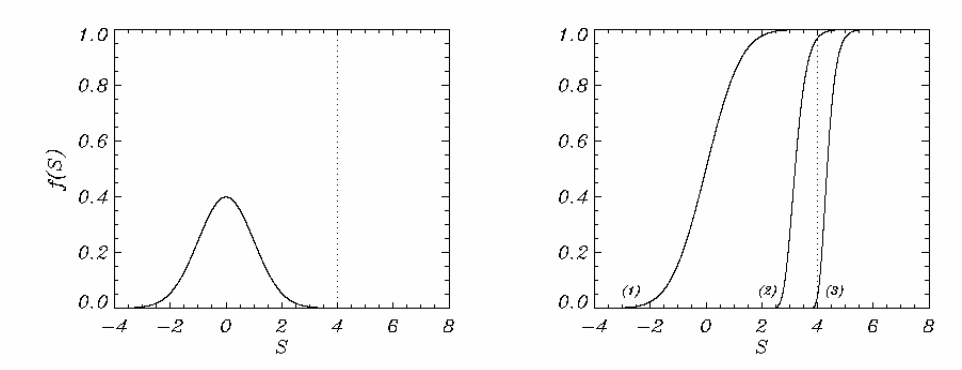

Figure 1

2.3. Critical Database Size

As shown in this section, the probability of finding a matching pair takes a simple form for values of C(S) near unity. A critical database size at which the probability of a sister being found as a match drops significantly from unity is then defined.

Consider values of C(S), written as 1– δ, where δ is < 0.01. The probability that a sister with score S is at the first position in a sorted list of N scores is

$$Prob(SisterRank = 1; N) = C(S)^N = (1-\delta)^N = e^{N\ln(1-\delta)} \approx e^{-N\delta} = e^{-N(1-C(s))}$$

where we replace ln(1– δ) by the first term of its Taylor series, which is a good approximation for small δ. Let us define the critical database size $N_{CRIT}(M;S)$ as the database size for which the probability that a score S is within the first M positions in the sorted score list drops to 1/e (= 0.37). From the equation above, the critical database size for M = 1 is

$$N_{CRIT}(1;S) = \frac{1}{1-C(S)}.$$

For other values of M, the critical database size cannot be derived analytically. However, it is found numerically that

$$N_{CRIT}(M;S) \approx \frac{M}{1-C(S)}.$$

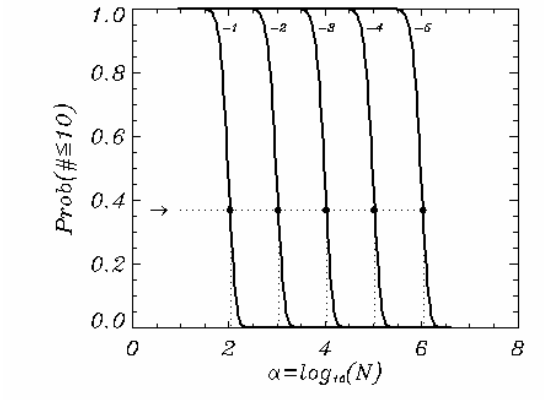

Figure 2

For example, Figure 2 displays the probability of a sister being found as a match (PROB(SisterRank≤10)) as a function of the database size for values of $\log_{10}$(1-C(S)) equal to –1, –2, –3, –4 and –5. All curves have the same shape, but with

different horizontal shifts (note the abscissa which goes like $\log_{10}$ of the database size). The probability drops to 1/e at the critical database size given by the above equation, with M = 10.

Computing the probability that a given sister with score S is found as a match for database sizes with up to one million exhibits requires a good knowledge of the cumulative distribution at least up to C(S)=0.99999, as found by replacing N_{CRIT} (10;S) by one million in the previous equation.

2.4. Computation of the Performance Curve

In the current study, the performance of the IBIS algorithms in a database of size N and for mark type T is defined as the proportion of sisters that are found with a score among the 10 largest scores, i.e., with a rank ≤10 in the sorted score list. If the performance is computed from actual scores, i.e., without any modeling, there is no probability involved.

$$P(SisterRank \le 10; N, T) = \frac{1}{N_A} \sum_A ([SisterRank \le 10; N, A\ , T] = true)$$

where the right term in the summation is Boolean, and where the summation is over all references A.

The generalization for a simulated non-sister database of size N consists in replacing the Boolean term by the probability that each sister is found within the first 10 positions in the sorted score list.

$$P(SisterRank \le 10; N, T) = \frac{1}{N_A} \sum_A Prob(SisterRank \le 10; N, A, T)$$

where the individual probabilities on the right are

$$Prob(SisterRank \le 10; N, A, T) = \sum_{M=0}^{9} \frac{N!}{M!(N-M)!} [1 - C(S_{AA'}; A, T)]^M \, C(S_{AA'}; A, T)^{N-M}.$$

Here, $S_{AA'}$ is the known score between reference A and its sister.

When several mark types T are combined (for example, the firing pin, breech face and ejector mark of a cartridge case), the overall performance is defined as the proportion of sisters that are found with a score among the 10 largest scores *for at least one of the mark types*. Again, if the performance is computed from actual scores, there is no probability involved.

$$P(SisterRank \le 10; N) = \frac{1}{N_A} \sum_A MAX_T ([SisterRank \le 10; N, A, T] = true)$$

The generalization is performed as before by replacing the Boolean term by probability

$$P(SisterRank \le 10; N) = \frac{1}{N_A} \sum_A MAX_T \left(P(SisterRank \le 10; N, A, T)\right).$$

In this work, the performance curve is defined as the functions P(SisterRank≤10;N, T) or P(SisterRank≤10;N), and is displayed as a Performance (P) versus $\log_{10}$(N) plot, as shown in Figure 3. The large black dot in Figure 3 indicates the measured performance as derived from the input scores.

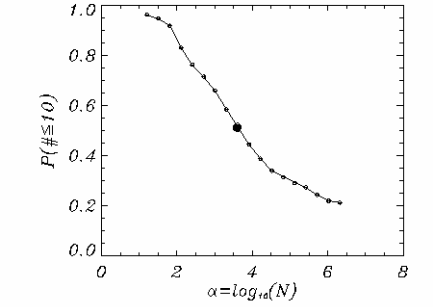

Figure 3

3. Determination of the Non-sister Score Distribution

In the previous section, the mathematical expressions for the relevant probabilities were derived by assuming a known non-sister score distribution C(S). However, this distribution is not known with infinite precision since the number of available scores is finite. It was also shown from the critical database size that the cumulative distribution must be known at least up to C(S) = 0.99999 to be able to apply the model for database size of one million. In this section, we cover the second phase in the development of the model, which is the determination of a valuable non-sister score distribution as derived from the available scores.

3.1. Missing Data Issue

All scores must be available to determine a complete "experimental" cumulative distribution. It turns out that not all scores are provided by IBIS.

Actually, to reduce computing time, the IBIS correlation process is divided into two steps. A crude correlation is first performed; its sole purpose is to efficiently segregate the exhibits that have low and medium/high similarity to the reference. A second, more sophisticated, correlation is performed on the promising candidates. While adjustable in principle, the amount of exhibits that are filtered in the second pass has been set to 20% in the current study; a value that reflects the conditions in the field. Therefore, the final correlation score is not computed for 80% of the sample.

Figure 4 (left) displays the non-sister score distribution for a 32 Auto cartridge case which was correlated against nearly 500 pairs of cartridge cases that were acquired at FT. IBIS returned 20% of the score; the remaining 800 score values were set to zero by definition. The resulting cumulative distribution is displayed in Figure 4 (middle). Another representation of the cumulative distribution, Figure 4 (right), enhances the behavior in the right wing of the cumulative distribution, and is more useful in the context of the current work; the ordinate has been changed to $\log_{10}(1\text{-}C)$.

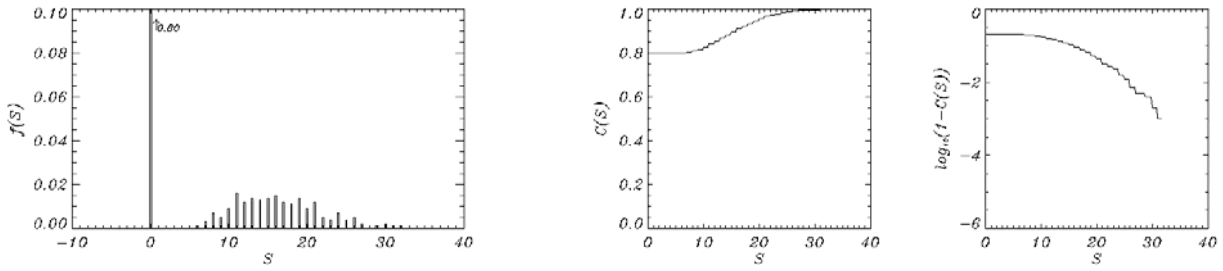

Figure 4

The last figure reveals the noisy character of the right wing of the cumulative distribution, which becomes undetermined for $\log_{10}(1\text{-}C) \leq -3$. Data is clearly lacking there, since only 200 scores are available. However, as explained in the previous section, knowledge of C up to 0.99999 (or equivalently, $\log_{10}(1\text{-}C) = -5$) is required to simulate the behavior of database with one million exhibits.

The strategy that we use to circumvent the missing data problem relies on three hypotheses:

The missing low scores (more than 80% of the data) have no impact on the projected performance in a large database.

A "universal" normalized distribution, which is not a function of the reference, can be built from all the available data. If true, 1000 references, each contributing 200 scores, would provide enough scores to build a smooth universal distribution.

The behavior of the cumulative distribution in the far right wing of the universal distribution is smooth enough to allow some extrapolation with simple curve-fitting techniques.

These hypotheses are described in the next three sections.

3.2. Missing Low Scores

As described in Section 2, the performance of the algorithm for a database of size N is the average of the probability that any sister is found as a match.

$$P(SisterRank \le 10; N, T) = \frac{1}{N_A} \sum_A Prob(SisterRank \le 10; N, A, T)$$

As indicated in Section 2, any sister pair with score $S_{AA'}$ does not contribute significantly to the performance curve at databases sizes N that are larger than the critical database size N_{CRIT} $(10;S_{AA'})$, if

$$N > \frac{10}{1 - C(S_{AA'})} \rightarrow C(S_{AA'}) < 1 - \frac{10}{N}.$$

In this study, we are interested in databases of size N that are larger than 1,000. For such databases, any sister pair with a score satisfying

$$C(S_{AA'}) < 1 - \frac{10}{N} < 0.99$$

does not contribute to the performance curve, i.e., they have a negligible probability of being in the first 10 positions. The detail of the cumulative distribution for $C(S_{AA'}) < 0.8$, which is lost because of the first pass in the correlation algorithm, has no effect on the performance curve. Simulations that were performed with arbitrary cumulative shapes for these score values yielded no change in the computed performance curve. The effect of the missing low scores is therefore negligible.

3.3. Computation of the Universal Normalized Cumulative Distribution

Figure 5 shows the experimental non-sister distribution for two reference exhibits of the same caliber, and with the same type of mark. The distributions are significantly different. No reasonable "universal" non-sister score distribution, independent of the reference exhibit, can be defined by averaging the experimental distributions.

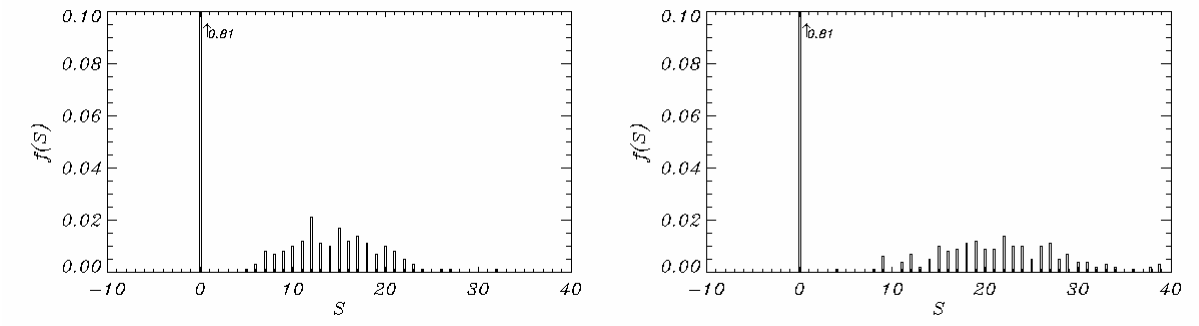

Figure 5

However, it turns out that the right wing of the non-sister score distributions that are associated with different references have a similar shape when they are properly normalized. The normalization process uses the average and variance of the *available* best scores (those provided by the second pass only, the other scores being set to zero). The score between reference A and any non-sister is normalized by the following shift and rescale operations,

$$S_N(A) = \frac{S - \mu(A)}{\sigma(A)}$$

where average μ and standard-deviation σ are computed for each reference A. Figure 6 shows the normalized distributions that correspond to the distributions that are shown in Figure 5.

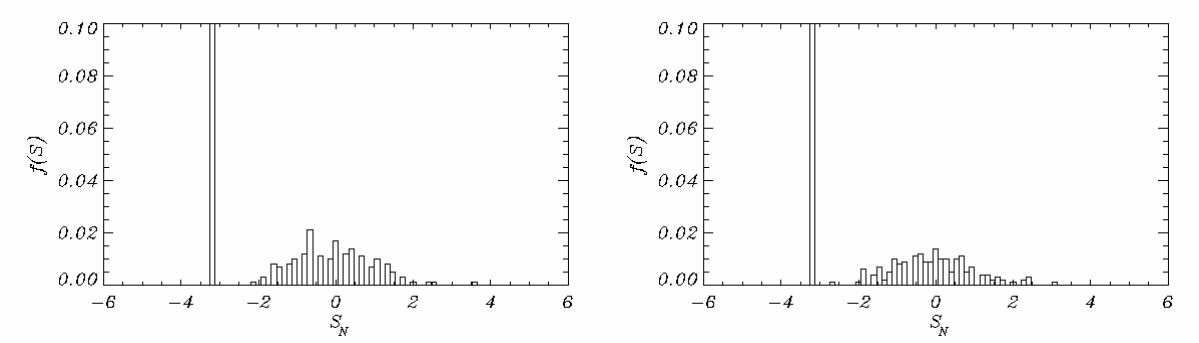

Figure 6

We define the "universal" normalized distribution as the average of the reference's experimental normalized distributions, for a given caliber and type of mark. The whole process is described below.

For each reference A, compute μ(A) and σ(A) by averaging all available non-sister scores S(i;A) from the second pass (i.e., excluding the missing null scores)

$$\mu(A) = \frac{1}{N'}\sum_{i=1}^{N'} S(i;A),$$

$$\sigma^2(A) = \frac{1}{N'-1}\sum_{i=1}^{N'} \left(S(i;A) - \mu(A)\right)^2$$

where N' is the number of available non-sister scores.

Define a set of predetermined normalized scores $\{S_{N,k}\}$, where k is between 1 and 1000. We use 1000 equidistant values between –20 and 20.

For each reference A, compute its normalized cumulative distribution over the set of predetermined normalized scores,

$$C(S_{N,k};A) = \frac{1}{N_S}\sum_{i=1}^{N_S}\left(\frac{S(i;A)-\mu(A)}{\sigma(A)} < S_{N,k}\right), \qquad 1 \le k \le 1000$$

where N_S is the total number of all null and non-null scores that are associated with reference A.

Compute the universal normalized distribution by averaging the normalized cumulative distributions that were computed above.

$$C_{UNIV}(S_{N,k}) = \frac{1}{N_A}\sum_{A} C(S_{N,k};A)$$

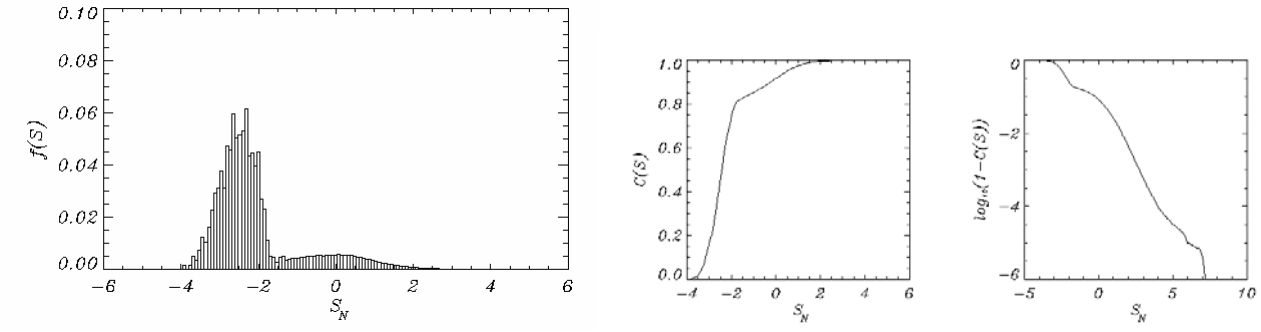

Figure 7

Figure 7 shows the universal normalized distribution for the 32 Auto caliber and its cumulative distribution. At values of S<–2, the behavior looks different than that of the individual normalized distributions (for example in Figure 5) since there is no high and thin peak anymore. This is because the null scores from IBIS do not contribute to a unique normalized score

since each reference has its own normalization coefficients. For a given reference A, the normalized score that is associated with the null scores is $-\mu(A)/\sigma(A)$, which differs from one reference to another. The high peaks, when averaged over all references, yield a wide peak in the universal normalized cumulative distribution between S= –4 and –2. The contribution of these values of S is not relevant for the performance curve anyway, as discussed in the previous section 3.2.

3.4. Extrapolation of the Right Wing of the Universal Normalized Cumulative Distribution

The predicting property of the model relies strongly on the treatment of the right wing of the cumulative distribution for $C(S_N)$ values near unity. Since the amount of data is limited in that region, the experimental normalized distribution must be replaced by an analytical function that fits the available data and has a reasonable behavior for larger values of S_N. To enhance this region, a fit is not performed on C, but rather over $\log_{10}(1\text{-}C)$.

Figure 8 shows the experimental normalized cumulative distribution that is associated with the score distribution of Figure 7.

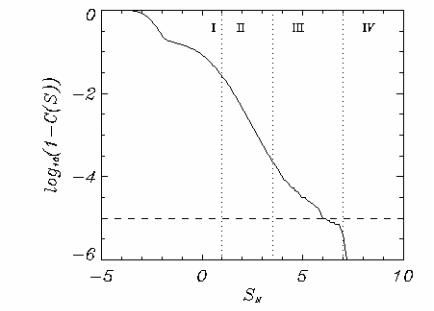

Figure 8

There are four regions in the curve (I–IV). For low values of S_N, the cumulative distribution is sensitive to the normalization of the null scores, which has no effect in the performance. In the second and third regions, the cumulative distribution looks like a decreasing function of the form x^{-n} but the slope changes between both regions. Sometimes, there is a sudden slope change for very high values of S (S > 6) due to missing data. The cumulative distribution cannot be used as such, since the detail of the distribution in the fourth region could have an impact on the performance curve for a database size that is close to one million. In this work, we assume that the slope of the cumulative distribution changes smoothly in the right wing, and we replace the cumulative distribution in the second and third regions by an analytical function which is extrapolated in the fourth region. Note that in Figures 8 and 9, the horizontal dotted line crosses the minimal values of $\log_{10}(1\text{-}C)$, for which knowledge is required to simulate the behavior of database sizes that are equal to one million.

The fitting procedure consists of fitting the cumulative distribution in regions II and III. Using an "eye" inspection, it has been found that S_N values between [1, 2.5] are within region II for all distributions that were studied. To compute the wing of the cumulative distribution for S_N values within that interval, we take a function of the form

$$F(x) = a + bx^n,$$

where b is assumed to be negative to ensure a decreasing function. The real coefficients a, b and n are found by using the method in Appendix A.

A similar fit is also performed for the third region in the S_N-interval [4, 7]. The final analytical solution is the maximum of both fitted functions. Figure 9 shows the fitting steps, with the resulting analytical fit. No case was found for which the experimental distribution in the third region is lower than the first fitted function that is extended to that region.

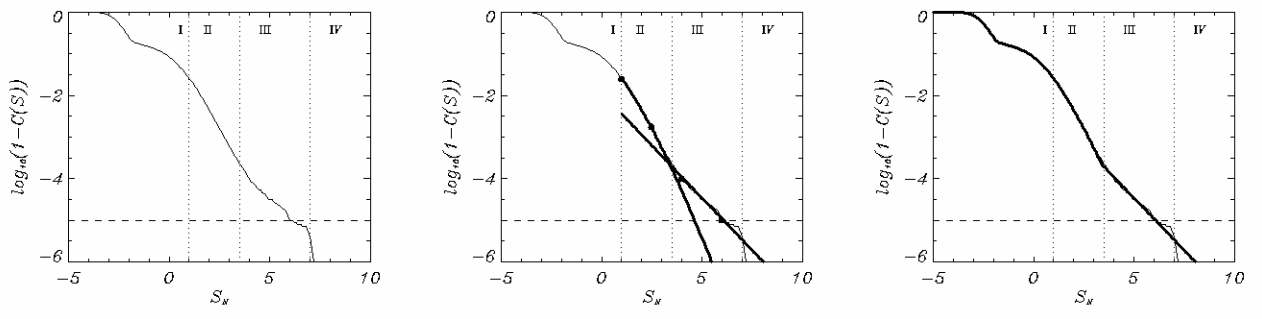

Figure 9

4. Error Analysis

This section provides an overview of the third and last phase in the development of the model, which is the error analysis. More details are provided in the Appendix B.

Three sources of errors were considered.

Uncertainty in the score normalization of sister pair scores.

As derived in previous sections, the probability that a A's sister (with score $S_{AA'}$) is found as a match is a function of $C(S_{AA'}; A)$, the A's cumulative distribution of non-sister scores at $S_{AA'}$. The value of the cumulative distribution is found in two steps, first, by normalizing the non-sister score with average and variance terms, and then by using the resulting score as an argument of the universal normalized cumulative distribution

$$C(S_{AA'};A) = C_{UNIV}\left(\frac{S_{AA'} - \mu(A)}{\sigma(A)}\right)$$

An error on the average and/or variance terms above would change the value of the cumulative distribution slightly, and change the probability that a sister with score $S_{AA'}$ is found as a match. As described in Appendix B, simulations have been performed in which the average/variance terms can vary within reasonable bounds. The errors that were deduced from the simulation were negligible. This uncertainty is not the main source of error.

Uncertainty of the non-sister score distribution, Method I.

The experimental distribution of non-sister scores is the result of randomly selecting N exhibits from a very large pool, which can be defined as the set of all possible cartridge cases that have the proper caliber. The real non-sister score distribution is not known; only the experimental one is known. As described in Appendix B, simulations have been performed in which N exhibits are randomly selected from an infinite pool that has a score distribution that is identical to the experimental distribution; the experimental universal normalized distribution is assumed to be close to the "real one". The "simulated" non-sister distribution that we obtain differs slightly from the distribution that we obtained with the real scores, having a measurable impact on the performance curve. This process was repeated several times to measure the variation of the performance curve. The result of this simulation yields error bars that were much smaller than those that were computed from the next source of errors.

Uncertainty of the non-sister score distribution, Method II.

The ideal way to measure the impact of the uncertainty of the non-sister score distribution on the performance curve is to acquire several new databases, each having 500 pairs of the same caliber, and compute the performance curve each time. Since this is unrealistic, we suggest the following approach. The available non-sister scores are split into a small number of K disjoint subsets. For each subset, the performance curve is computed by finding the probability that each sister (among the 500 pairs) is found as a match when the non-sister score distribution is determined from the subset's distribution. Here we assume that the K distributions differ between themselves in the same way that the original experimental distribution differs from the real one.

The non-sister distribution differs slightly between subsets, which leads to measurable differences on the performance curve. This was found to be the larger source of error. We therefore adopted this method to compute the error bars for the performance curves that will be presented in the Section 5.

5. Experimental Results

5.1. Input Data

Table 1 provides an overview of the acquisitions that were performed at FT. For four calibers, nearly 500 sisters, mostly pairs from the same ammunition type, were randomly selected from samples that were provided by the Pittsburgh State Police department, and acquired with IBIS. They have been correlated against themselves and against a larger database, which includes additional data from IBIS users in other areas such as Germany, South Africa, and New York.

Table 1 lists the measured performance for these calibers against two databases, the "small" database, which is the set of acquired pairs, and the larger database. For 9mm, 32 Auto and 45 Auto, breech face marks and firing pin were acquired over the 500 pairs. Similarly, for 9mm, breech face marks, firing pin, and ejector marks were acquired for 78 pairs. The rimfire firing pin mark is relevant for 22 caliber only.

Caliber	**Mark**	**FT Acquisition**	**Database Size**	**Performance (%) BF/FP Combined Marks**
9mm	BF/FP	434 pairs	868	53, 74, 84
			56000	39, 53, 66
32 Auto	BF/FP	500 pairs	1000	35, 84, 87
			10700	25, 72, 76
45 Auto	BF/FP	474 pairs	948	55, 57, 73
			3535	47, 49, 65

Caliber	**Mark**	**FT Acquisition**	**Database Size**	**Performance (%) BF/FP/EJ Combined Marks**
9mm	BF/FP/EJ	78 pairs	4030	51, 56, 46, 94

Caliber	**Mark**	**FT Acquisition**	**Database Size**	**Performance (%)**
22	Rimfire firing pin	500 pairs	1000	87
			3070	87

Table 1

This table can be summarized as follows:

For all four calibers, the combined performance (first 10 positions for the breech face, firing pin or rimfire firing pin marks, if applicable) is in the 73–87% range for a database with 1000 exhibits.
All (except 22 caliber): All three marks (breech face/firing pin/ejector), when available, provide independent information that should be combined when searching for a match.
32 Auto: For all marks, there is a 10% decrease in performance when increasing from a 1000-exhibit to a 10000-exhibit database.

9mm: For all marks, there is a 15%–20% decrease in performance when increasing from a 1000-exhibit to a 56000-exhibit database.
45 Auto: For all marks, there is a 10% decrease in performance when increasing from a 1000-exhibit to a 3000-exhibit database.
22: For the rimfire firing pin mark, there is a less than 1% decrease in the performance when increasing from a 1000-exhibit to a 3000-exhibit database.

5.2. Computation of the Performance Curves for Breech Face and Firing PinMarks

The simulated performance curve has been computed for three calibers: 9mm, 32 Auto and 45 Auto. Figure 10 shows their respective behavior with error bars.

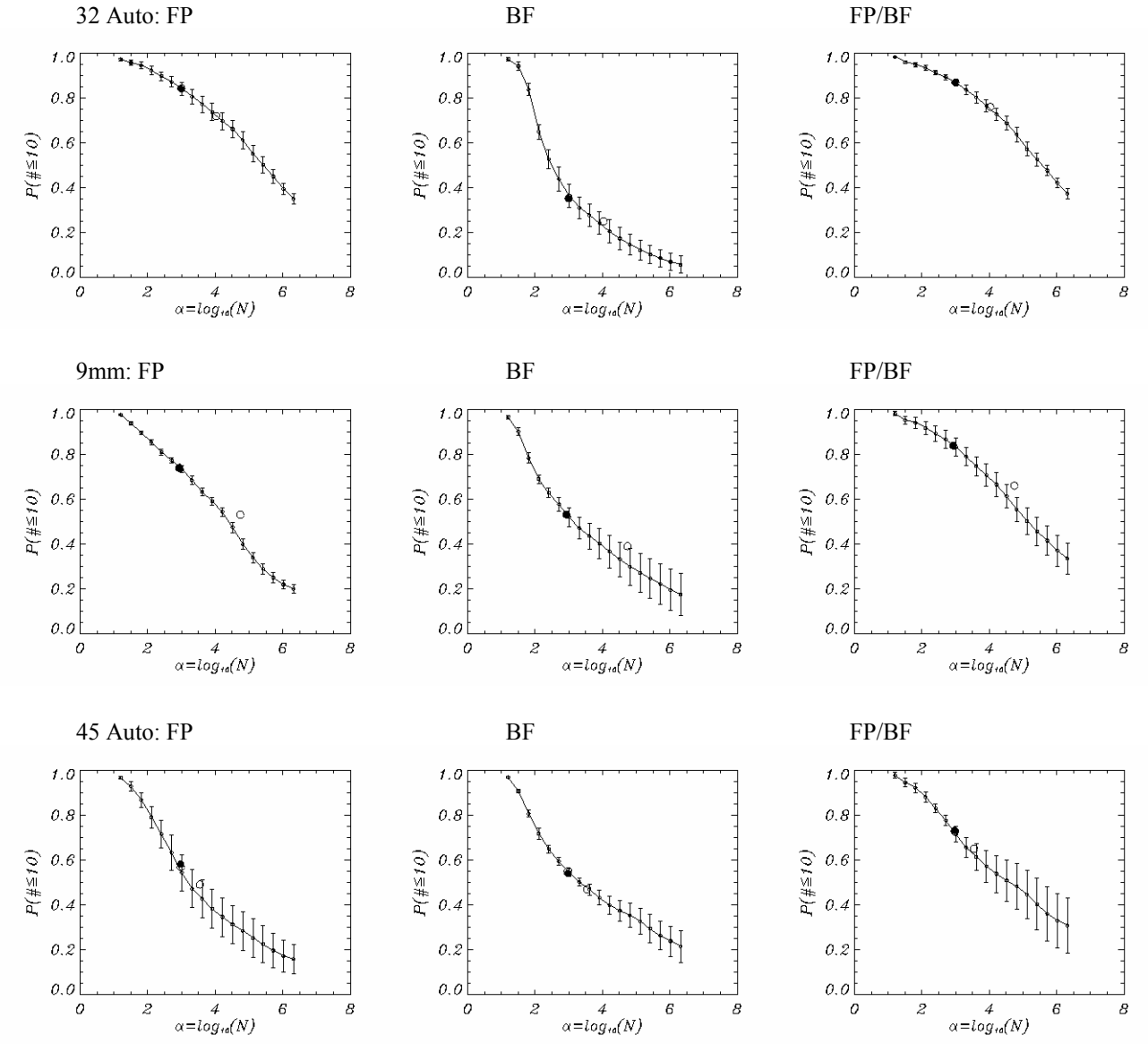

Figure 10

Figure 10 can be summarized as follows:

A large black dark dot indicates the measured performance of the set that is used to train the model (about 500 sister pairs). The agreement with the simulated performance curve for the same database size is very good. We must stress that the measured performance of the set and the predicted performance are computed by different processes. The measured performance is the actual proportion of references with sister scores that are within the first 10 positions; the simulated performance curve is computed from the probability that the sister's score is within the first 10 positions, which is found from a universal normalized distribution, independent of the reference, and from the average and variance of the non-sister score distribution of each reference.

A large white dot indicates the measured performance when the reference set (500 sister pairs) is correlated against the large database (4000 to 56,000 exhibits). Agreement is in general very good, the only exception being the 9mm firing pin, to be discussed below. The combined (breech face and firing pin) performance lies within the error bars for that caliber.

As expected, the performance decreases with database size. The relative performance of the breech face and firing pin acquisitions differs from one caliber to another. However, for all three calibers, the expected combined performance is close to 30–35% for a one million-exhibit database when looking at the top 10 ranking samples.

For the 9mm firing pin, there are at least two explanations for the slight discrepancy between the actual performance of the large database (large white dot) and the performance curve (large black dot).

Some sources of error, that could be extracted from the input data set (500 pairs), either remains to be found or have not been modeled properly in Section 4.

A significant source of error arises from the different statistical properties between the training set (500 pairs) and the large database (56,000 exhibits). This is plausible since the 500 pairs have been provided by Pittsburgh State Police and could have class characteristics that are consistent with the ammunition types and/or firearm models that are available in that region, while the large database data is mostly from Germany and South Africa with, most likely, different class characteristics. In this case, the non-sister scores among the 500 pairs should be higher (since they would result from correlations between cartridge cases that share similar class characteristics), and this, in turn, would imply a lower probability of finding matches.

These explanations however, do not clarify why there is a discrepancy for 9mm, but not for 32 Auto. One possibility is that, since the 9mm is the most popular caliber, a greater number of manufacturing processes are available for that caliber. It is more probable that the class characteristics vary from one geographical region to another.

A digression about the error bars is necessary at this point. Section 4 and Appendix B explain that the error bars are computed by generating performance curves from five different subsets of the non-sister scores that are used to train the model. However, using the statistical model requires a minimal amount of data. It turns out that 500 pairs was not enough to compute valuable error bars, since only 20% of the scores are saved in the database, out of which five disjoint subsets must be created. The solution was to modify IBIS slightly by removing the first pass in order to work with larger subsets of scores. The error bars were then computed from these. Note that the performance on IBIS, as measured by the number of found matches in the training set, remained the same whether the first pass was applied or not.

Another valuable piece of information is the expected change in the performance curve if the definition of a match is changed. Until now, a sister was defined as a "found match" if its score is among the M = 10 largest scores. Figure 11 shows the behavior of the performance curve for the combined marks (breech face/firing pin), for M = 1, 10, 25, 100 and 250 (the performance associated with M = 10 is shown in bold).

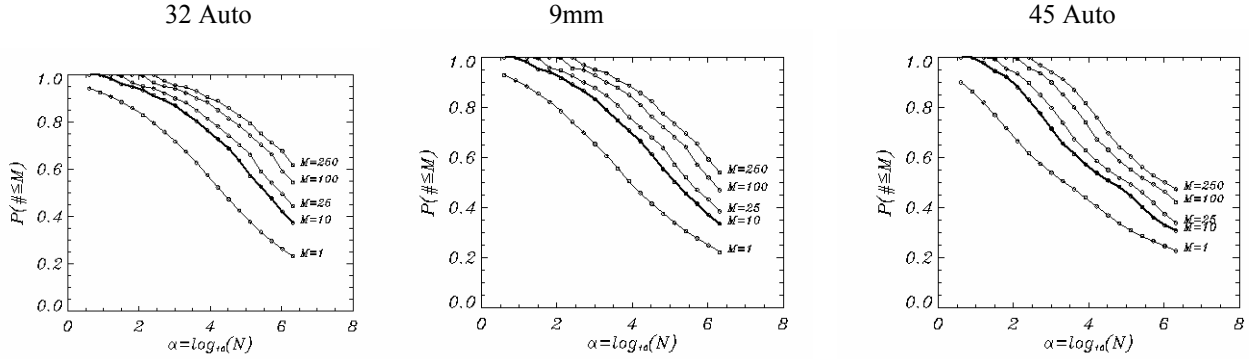

Figure 11

The three curves share a common behavior. The performance curve has a significant jump between the first and the tenth positions, and then increases more and more slowly as M increases. For example, the gain in performance between the first 10 and the first 100 positions is similar to the gain in performance between the first position and the tenth position.

5.3. Computation of the Performance Curves for Ejector and Rimfire Firing Pin Marks

For the smaller set of 9mm with all three marks acquired (78 pairs) and for 22 caliber, we did not correlate with the second version of IBIS (i.e., without first pass); therefore, we do not have error bars for them. Figure 12 shows the performance curves for these two calibers with no error bar. In both cases, the agreement between the experiment and the model remains excellent. Furthermore,

For 9mm, the combined performance is still about 50% in a one million-exhibit database. Each individual mark's performance drops to 10–25%, but their contribution almost adds up, thus confirming the benefit of using complementary information from the different types of marks.

For 22 caliber, there is only one mark to correlate. Its performance reaches 45% for a one million-exhibit database.

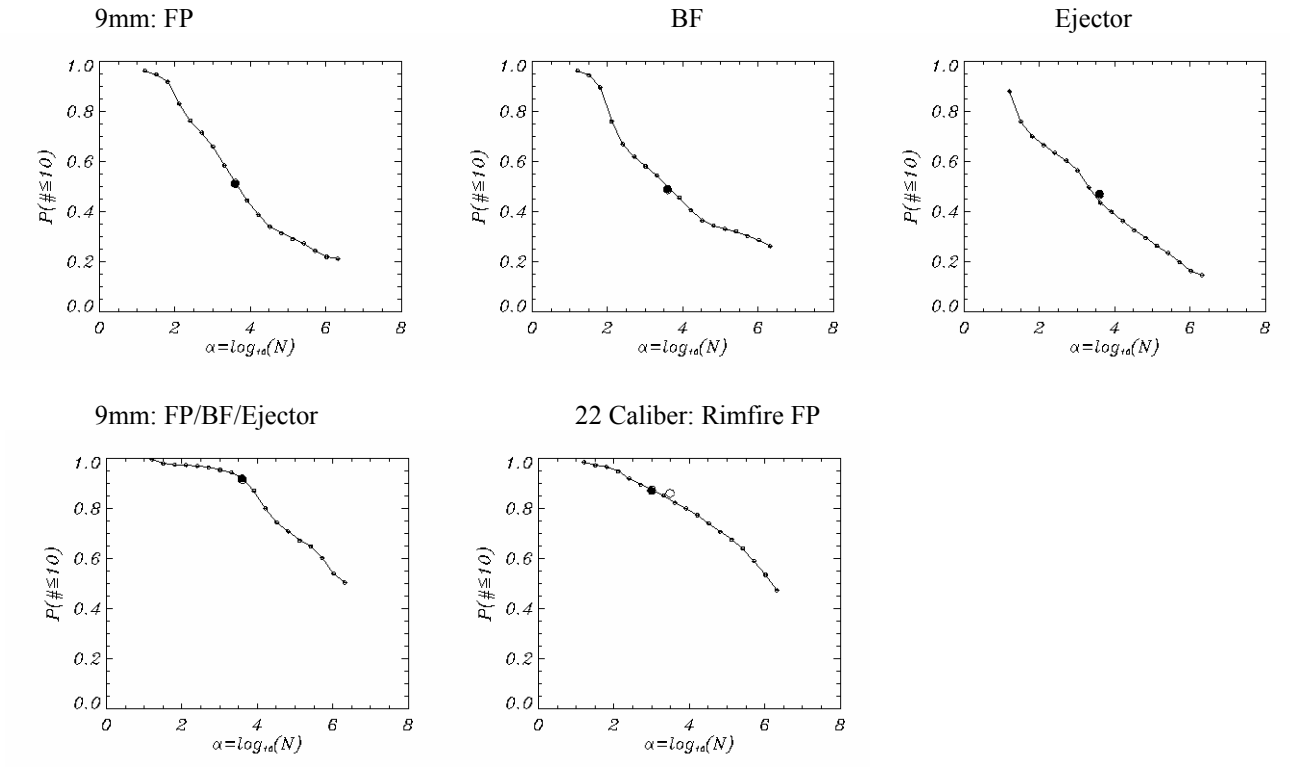

Figure 12

5.4. Extrapolation from the Large Database Data

From the agreement found between the measured performance of the large database and the projected performance of the model (refer to the large white dots in Figures 11 and 12), we can safely use the model to extrapolate from the performance found experimentally in the available large database.

In the previous sections, the model was trained with the non-sister scores that were found by correlating each acquired cartridge case among 500 sister pairs against the other ones in the same set. In this section, the model is trained with the non-sister scores that were found by correlating the 500 sister pairs with a larger database. This is done for 32 Auto and 9mm since the size of their large database (10700 and 56000 respectively) is significantly larger than the size of the first

training set (nearly 500 pairs). There is no point to extrapolate for 45 Auto and 22 caliber since their large database was only three or four times larger than the training set.

For 32 Auto, it turns out that the performance curve that was computed from the large database data, and the one computed from the original smaller training set, are almost identical. Figure 13 shows the situation for 9mm. Here, the model predicts better performance with the large database data. This is consistent with the slight discrepancy that was already discussed.

Large Database (9mm): FP BF FP/BF

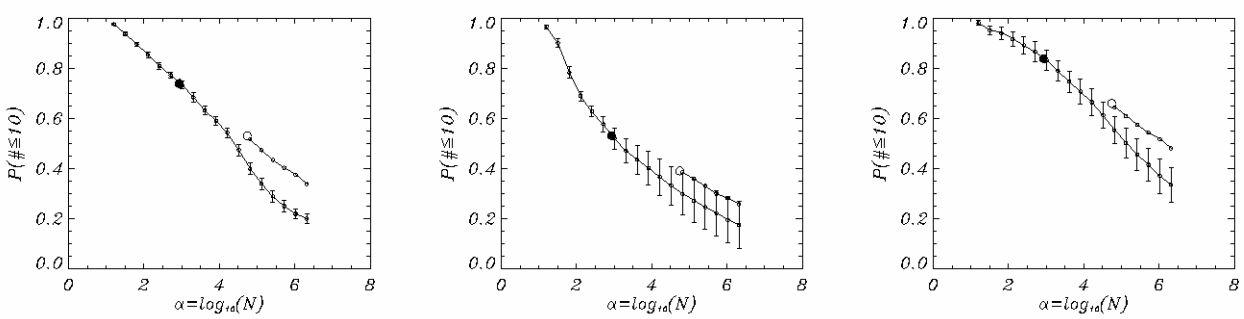

Figure 13

6. Conclusion

A statistical model has been developed to predict the performance of the IBIS correlation algorithm as a function of the database size. From a set of known sister and non-sister pair scores, the model is used to predict the proportion of *those* sister pairs that would be found as a match as a function of the database size. The conclusion inferred from the model should be applied only for the sister pairs that are used as input. It is also assumed that the non-sister score distribution in the *simulated* database is the same as in the *experimental* database.

The model uses the fact that the non-sister score distribution between a given reference and the rest of the database has a "universal" shape, independent of the reference itself, if the scores are properly normalized. A smooth statistical distribution is generated by combining the normalized non-sister scores, and by extrapolating the available data in the right wing by an analytical function.

The model has been trained using real data. Nearly 500 pairs of cartridge cases, with each pair fired from the same firearm, have been acquired with IBIS, for four different calibers (9mm, 32 Auto, 45 Auto and 22). The acquired marks were the breech face mark, ejector mark, and firing pin mark for the first three calibers and the rimfire firing pin mark for the last caliber. Furthermore, the ejector mark has been acquired for 78 pairs of 9mm cartridge case. Each acquired cartridge case was correlated against the rest of the database. A performance curve, that predicts the performance of the IBIS system as a function of the database size, was computed for the four calibers from the sets of correlation scores. The expected performance decreases from 80% to 30–45% when the database size increases from 1000 to one million exhibits. The model also confirms that acquiring all available marks improves the combined performance. The breech face, firing pin and ejector marks thus provide complementary information.

To put this in perspective of an operational environment, a database of a million images for a correlation could be significantly reduced through segmentation of the database. Currently some crime system programs have database sizes in the 100's of thousands and one program has even reached over a million exhibits for the total database. The largest correlation requests, for the most common calibers, are on average no larger than 10,000 samples in each request. These databases are segmented based on geography, and some basic type determination characteristics of the samples. These techniques could be expanded further with other segmentation parameters in order to minimize samples sets during a correlation. Some preliminary techniques are outlined in detail in the FT White Paper "Tool Mark Imaging Reference Files".

Another point to put in perspective is that the results and information in this model are related to current operational procedures for reviewing of images. Currently an IBIS user would look at the top M = 10 samples for each image type. The

users use the IBIS Multi-viewer which is a tile screen that allows users to view multiple images from a correlation request on a single screen. As was shown in Figure 11 looking at more results (M = 10, 25, 100, 250) also increase the probability of finding a match. The user can easily configure the tile screen to display more images so that reviewing procedures can be optimized. A user could in effect visualize a single tile screen of 49 images (7 x 7) very quickly, each subsequent group of images would be a mouse click away.

7. APPENDIX A: Fitting Procedure

This appendix presents the method that is used to fit the wing of the cumulative distribution for S_N values in that interval. We use a function of the form

$$F(x) = a + bx^n,$$

where b is assumed to be negative to ensure a decreasing function.

The real coefficients a, b and n are found by minimizing the nonlinear chi-square function

$$\chi^2 = \sum_i (a + bx_i^{\ n} - y_i)^2.$$

The coefficients are found in two steps. First, by forcing the partial derivatives with a and b to be zero, we obtain two linear conditions on a and b that minimize the function for a given exponent n.

$$a = \frac{\begin{vmatrix} \sum y_i & \sum x_i^n \\ \sum y_i x_i^n & \sum x_i^{2n} \end{vmatrix}}{\begin{vmatrix} \sum 1 & \sum x_i^n \\ \sum x_i^n & \sum x_i^{2n} \end{vmatrix}}, b = \frac{\begin{vmatrix} \sum 1 & \sum y_i \\ \sum x_i^n & \sum y_i x_i^n \end{vmatrix}}{\begin{vmatrix} \sum 1 & \sum x_i^n \\ \sum x_i^n & \sum x_i^{2n} \end{vmatrix}}$$

The problem now becomes one dimensional, but remains nonlinear. A brute force is then performed over values of n (while changing a and b accordingly using the above equations) to find the global minimum of the chi-square function.

8. APPENDIX B: Error analysis

8.1. Normalization error

As derived in Sections 2 and 3, the probability that a A's sister (with score S) is found as a match is a function of A's cumulative distribution of non-sister scores at S, that is, C(S; A). The value of the cumulative distribution is found in two steps, first, by normalizing the non-sister score with average and variance terms, and then by using the resulting score as an argument of the universal normalized cumulative distribution,

$$C(S_N; A) = C_{\ NORM}\left(\frac{S - \mu(A)}{\sigma(A)}\right)$$

An error on the average and/or variance terms would impact the value of the cumulative distribution, and the probability to be found as a match.

To evaluate this effect on the performance curve, we used the two following properties of the average/variance of a set of independent points, which are formally true for gaussian distribution of points, but still approximately hold if the number of points is much greater than 30 (Schaum, probability and statistics).

The distribution of the computed average μ(A) follows a normal distribution with average μ_T(A) and variance $\sigma_T(A)/\sqrt{N}$, where the T subscript stands for the "true" value.

The distribution of the computed variance follows a chi-square distribution with an n-1 degree of freedom, with average n-1 and variance 2(n-1).

$$\mu(A) \to N(\mu_T(A), \frac{c(A)}{\sqrt{n}})$$

$$\frac{n\sigma^2(A)}{\sigma_T^2(A)} \to \chi^2(n-1)$$

The following simulation was then performed:

For each reference A, we generated new μ(A) and σ(A) values, of the form

$$\mu_{RAND}(A) = \mu_T(A) + X\frac{c(A)}{\sqrt{n}}$$

$$\sigma^2{}_{RAND}(A) = \frac{\sigma^2(A)}{n}\left[(n-1) + Y\sqrt{2(n-1)}\right]$$

where X and Y are generated randomly between [-2,2]. The random generator is called for each reference.

The performance curve was computed with the resulting new cumulative distribution $C_{RAND}(S;A)$ for each reference A.

A set of performance curves was computed by repeating this process 100 times.

For each database size, the performance error was computed as the standard deviation of the set of performances.

The result of this simulation yields error bars that were negligible and essentially undetectable on the performance curve. The uncertainty on the normalization is not the main source of error.

8.2. Theoretical Variability of the Performance Curve from the Variability of the Experimental Non-sister Score Distribution

The experimental distribution of non-sister scores is the result of randomly selecting N exhibits from a very large pool, which can be defined as the set of all possible cartridge cases that have the proper caliber. The real non-sister score distribution is not known, only the experimental distribution is known. In this section, simulations are performed for each reference A, in which we randomly select N non-sister scores from an infinite pool that has a score distribution that is identical to the experimental distribution (i.e., our experimental distribution for reference A is assumed to be close to the "real one"). The A's "simulated" non-sister distribution that we obtain differs slightly from the distribution of the real scores. As a consequence, the universal normalized distribution also differs from the one that we obtain from the real scores, with a measurable impact on the performance curve. If we repeat this process many times, and randomly select a new set of N non-sister scores, the distribution of non-sister scores would change slightly each time for every reference. The computed performance curve would change accordingly.

The variability of the cumulative distribution in the simulation is determined as follows. Let us consider the following random experiment: a score is computed by correlating a given reference A with an acquired non-sister cartridge case. Since IBIS has a finite number J of possible output scores[3], the output of the experiment can be classified in J mutually exclusive outcomes with probability p_i (i = 1 to J), respectively. The i^{th} possible outcome is defined as the score S_i within the list of allowed scores from IBIS, which is assumed to be sorted from the lowest to the highest. Furthermore, let N_K be the random variable that measures the number of scores that are less than a given threshold score S_K, with $1 \le K \le J$, from a set of N independent scores. Since a given score can only be less or not less than S_K, the relevant distribution of N_K is the binomial distribution

[3] For example, in the case of the breech face/firing pin, the possible scores are integers between 0 and 800.

$$B(N_K) = \frac{N!}{N_K!(N-N_K)!} P_k^{N_K} (1-P_K)^{N-N_K}$$

$$P_K = \sum_{i=1}^{K-1} p_i$$

with average NP_K and variance NP_K (1- P_K). Our goal is to now compute the variance on the cumulative distribution $C(S_K)$

$$\sigma^2(C(S_K)) = \sigma^2(\frac{N_K}{N}) = \frac{1}{N^2}\sigma^2(N_K) = \frac{P_K(1-P_K)}{N} \approx \frac{C(S_K)(1-C(S_K)}{N},$$

where the parameter P_K is replaced by its experimental value $C(S_K)$.

The effect of this variability on the performance was evaluated by the following simulation.

For each reference A, we generated a new experimental cumulative distribution of the form

$$C_{RAND}(S;A) = C(S;A) + X\sqrt{\frac{C(S;A)(1-C(S;A))}{N}}$$

where X is generated randomly between [-2,2]. The random generator is called for each reference.

The universal normalized cumulative distribution was computed by averaging over the new reference's cumulative C(S;A).

The performance curve was computed with the new universal normalized cumulative distribution.

A whole set of performance curves was computed by repeating this process 100 times.

For each database size, the performance error was computed as the standard deviation of the set of performances.

The result of this simulation yields error bars that were much smaller than those that were computed from the next source of errors.

8.3. Experimental Variability of the Performance Curve from the Variability of the Experimental Non-sister Score Distribution

As discussed in the previous section, if we could randomly select N exhibits from a very large pool many times, the histogram of non-sister scores would change slightly each time for every reference, and the computed performance curve would also change. In this section, the variation of the performance curve around some "average performance curve" will be also evaluated by generating new cumulative distributions. There is, however, no random process involved here.

Our strategy consisted of dividing the database into K disjoint subsets. The following was then performed for each subset:

For each reference A, we generated the experimental cumulative distribution from the non-sister scores using only the current subset of exhibits

The universal normalized cumulative distribution was computed by averaging over the reference's cumulative C(S;A, Subset).

The performance curve was computed with this new universal normalized cumulative distribution using all references A.

At this point, K performance curves have been generated, similar to those in Figure 14. The variability among the curves is expected to reflect the inherent variability between the disjoint subsets.

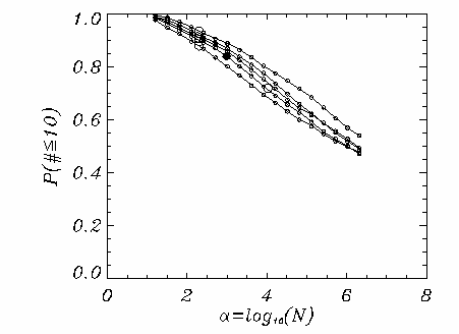

Figure 14

The last step was as follows:

> For each database size, the performance error was computed as the standard deviation of the set of performance curves.

This method was used with K = 5 disjoint subsets. This value of K was a good compromise between having a high subset size and a high number of subsets. It turns out that the performance error that was computed using this method is much larger than the performance error that was computed using the strategies that were presented in the two previous sections.

Here, we therefore assume, that the five distributions differ from each other in the same way that our original experimental distribution differs from the real one.