

Predicting large population data cumulative match characteristic performance from small population data

Amos Y. Johnson, Jie Sun, and Aaron F. Bobick

GVU Center / College of Computing
Georgia Tech, Atlanta, GA 30332
{amos, sun, afb}@cc.gatech.edu

Abstract. Given a biometric feature-space, in this paper we present a method to predict cumulative match characteristic (CMC) curve performance for large populations of individuals using a significantly smaller population to make the prediction. This is achieved by mathematically modelling the CMC curve. For a given biometric technique that extracts measurements of individuals to be used for identification, the CMC curve shows the probability of recognizing that individual within a database of measurements that are extracted from multiple individuals. As the number of individuals in the database increase, the probabilities displayed on the CMC curve decrease, which indicates the decreasing ability of the biometric technique to recognize individuals. Our mathematical model replicates this effect, and allows us to predict the identification performance of a technique as more individuals are added without physically needing to extract measurements from more individuals.

1 Introduction

For a particular biometric technique, identifying a person can be split into two problems: verification and recognition [4]. Verification implies that the system is trying to confirm that *you are who you say you are*, and recognition refers to a system that is trying to determine *who you are*. The verification problem is considered to be the easier of the two problems. A verification system has only to take a sample measurement (a measurable feature of identity) of the subject and compare it against the known identity of the person they are claiming to be. This is an one-to-one matching problem; whereas, the recognition task is an one-to-many matching problem, which is a much harder problem in comparison. A recognition system must compare the sample measurement against an entire database of known identities, to determine who the sample belongs to.

In this paper, we are concerned with evaluating the ability of a biometric technique to recognize individuals. To evaluate performance in the recognition paradigm, a cumulative match characteristic (CMC) curve [6] is employed. A CMC curve shows various probabilities of recognizing an individual depending on how similar their biometric measurements are to other individuals' biometric

measurements. The goal of this paper is to mathematically model the CMC curve for a feature-space of biometric measurements, and then use that model (using a small subset of the feature-space to estimate the parameters of the model) to predict CMC curves for larger groups of individuals without actually needing more individuals to calculate the results.

In the following sections, we explain the meaning of a CMC curve, and derive a mathematical model of the CMC curve. Finally, we show experimentally that this model can generate a CMC curve, and is able to predict the recognition results as the number of individuals in increases.

2 Meaning of a CMC Curve

Before we discuss what a CMC curve is, we first introduce the notion of gallery and probe sets. A gallery is a set of ID templates (in the form of feature vectors/data points), representing individuals in a database. Each individual has only one data point in the gallery, so if there are N individuals, then the size of the gallery is N . These gallery data-points are used to recognize unknown feature-vectors, where the unknown feature vectors come from a probe set. The feature vectors in the probe set have a corresponding ID template in the gallery, and the probe set may contain more than N data points¹.

In general, a CMC curve is transparent to the underlying feature space of measurements and the technique used to compute the similarity between measurements. However, this work is predicated on knowing the feature space and computing similarity between measurements using a L2norm or Mahalanobis distance. With this in mind, to create a CMC curve, the distance (L2norm or Mahalanobis) between the probe data-points and the gallery data-points are computed. Second, for each probe data-point a rank ordering of all the data-points in the gallery is compiled, from the data-point with the closest distance to the furthest distance. Last, the percentage of time a probe data-point's corresponding gallery data-point is within Rank n (between 1 and n , where $n \in [1, N]$) is computed.

A point to note about a CMC curve is that as the number of subjects increases, the percentage of time a probe vector's corresponding gallery vector is between 1 and n is lower as N increases. For example, a CMC curve with 50 individuals may have a curve with the rank 1 percentage being 81%; however, if N increases to 100 individuals, then the CMC curve rank 1 percentage is 71%². This means that as the number of individuals to be recognize increases the ability to exactly recognize them decrease. With this in mind, for a given biometric technique with any number of subjects, the CMC curve is always lower with a greater number of subjects.

¹ This simply means that each individual can have more than one data point in the probe set.

² These results were generated using synthetic data with the following parameters: dimension = 1, $\sigma_p^2 = 10000$, & $\sigma_i^2 = 1$

3 Mathematical Model of a CMC Curve

To model the CMC curve mathematically, we begin with some terminology. A target is an individual that we are trying to recognize, and an imposter is any individual that is not the target. Also, the term (target’s or imposter’s) template refers to a data point in the gallery set, and the term (target’s or imposter’s) measurement refers to a data point in the probe set.

We now define the random variable v_s to be the distance between a target’s measurement and its template \mathbf{q}_s , with density $p_s(v)$. We also define another random variable v_n^q to be the distance between a target’s measurement and an imposter’s template \mathbf{q}_n , with density $p_n^q(v)$.

We assume that a target’s or an imposter’s template $\mathbf{q}_{s/n}$ was generated from a d -dimensional Gaussian-density, referred to as the population density. The single Gaussian assumption is made for simplicity; however, this assumption may be replaced with a mixture of Gaussians, parzen windows, or an appropriate density that fits the actual data. With the appropriate density, similar steps to ones presented in this work may be followed to model the CMC curve.

Next, the measurement variation of the template is modelled as a d -dimensional Gaussian-density, $p_i(\mathbf{x}) = N(\mathbf{q}_{s/n}, \Sigma_i)$ (data points from this density are in the probe set), referred to as the individual variation (or individual density). Also, a pivotal assumption we make about the densities is that the individual variation is much lower than the population variation.

In order to simplify the model, we make the requirement that the individual density, $p_i(\mathbf{x}) = N(\mathbf{q}_{s/n}, \Sigma_i)$, has a spherical covariance-matrix with variance σ_i in each dimension. If Σ_i is not spherical, then each population (gallery) feature vector needs to be scaled per dimension by the corresponding standard deviation from Σ_i , each dimension in Σ_i must be normalized to unity, and the off diagonal terms of Σ_i are set to zero³. Also, a new Σ_p needs to be computed for the population distribution.

Next, since p_i is a zero-mean Gaussian, the distance density $p_s(v)$ can be represented by the square root of the χ^2 (chi-square) density, with normalized density parameter $\chi^2 = \sum_{j=1}^d \frac{x_j^2}{\sigma_i^2}$ [3], and $v_s = \sqrt{\chi^2}$, resulting in

$$p_s(v) = \frac{v^{(d-1)}}{\sigma_i^d 2^{(d/2-1)} \Gamma(d/2)} e^{-\frac{v^2}{2\sigma_i^2}}, \quad (1)$$

where the symbol Γ in the denominator is the gamma function.

We now need to derive the probability that a target’s measurement is within some distance k of an imposter’s template \mathbf{q}_n , namely $\int_0^k p_n^q(v) dv$. Since the individual variation is much lower than the population variation, this probability is merely the volume of a d -dimensional hypersphere with radius k times the

³ Setting the off diagonal terms to zero is plausible because we assume that the individual variation is due to measurement noise, so it should be independent.

probability $p_p(\mathbf{q}_n)$:

$$A(\mathbf{q}_n, k) = \int_0^k p_n^q(v) dv = p_p(\mathbf{q}_n) \cdot V_d \cdot k^d, \quad (2)$$

where

$$\begin{aligned} V_d &= \pi^{d/2} / (d/2)! \quad d \text{ even} \\ &= 2^d \pi^{(d-1)/2} \cdot \frac{((d-1)/2)!}{d!} \quad d \text{ odd.} \end{aligned}$$

If we compute the distance between a target's measurement and all templates and rank the templates by their distance, (given that a target's measurement is a distance k away from its template \mathbf{q}_s) we define the probability that the target's template is exactly rank n to be a binomial probability distribution [5],

$$B(R, \mathbf{q}_*, k) = C_{R-1}^{N-1} (A^*)^{R-1} (1 - A^*)^{N-R}, \quad (3)$$

where

$$C_{R-1}^{N-1} = \frac{(N-1)!}{(R-1)!(N-R)!}.$$

R is rank, N is the number of subjects, and A^* is probability that the a target's measurement is less than or equal to a distance k from an imposter's ID template.

Since the probability of A^* is not exactly known, we approximate it with the probability that a target's measurement is within some distance k of a *particular* imposter's template \mathbf{q}_n (see Equation 2), and Equation 3 becomes

$$B(R, \mathbf{q}_n, k) = C_{R-1}^{N-1} A(\mathbf{q}_n, k)^{R-1} (1 - A(\mathbf{q}_n, k))^{N-R}. \quad (4)$$

Next we integrate over all possible distances k , which a target's measurement may be at from its template

$$\int_0^\infty B(R, \mathbf{q}_n, k) p_s(k) dk,$$

and since we approximated A^* for a *particular* imposter's template \mathbf{q}_n , we integrate over all possible imposter's templates

$$CMC^*(R = n) = \int_{-\infty}^{+\infty} p_p(\mathbf{q}_n) \int_0^\infty B(R, \mathbf{q}_n, k) p_s(k) dk dq_n. \quad (5)$$

Finally, the probability that a probe data-point's corresponding gallery data-point is within rank n is

$$CMC(R) = P(R \leq n) = \sum_{i=1}^n CMC^*(R = i). \quad (6)$$

To solve Equation 5, numerical integration techniques [2] are used. We note that Equation 5 has a multiple integral over the population density. As the dimensionality of the population increases, the number of integrations also increases. As the number of integrations increases, the computational complexity of integrating Equation 5 increase. In the Appendix, we derive a simplified version of Equation 5 that approximates the equation and eliminates the multiple integral term. We note that using the simplified version slightly under estimates the CMC curve at the lower ranks and slightly over estimates at the higher ranks. In the following experiments, the first more accurate formalization of the CMC model is used, unless otherwise stated.

4 Experimental Evaluation

Finally, we demonstrate the ability of Equation 6 to predict CMC curves. We first demonstrate its use on synthetically generated data, and secondly on actual data from our gait-recognition technique [1].

4.1 Synthetic Data

For the synthetic data, we generate a gallery of templates from a population Gaussian distribution, $p_p(\mathbf{x}) = N(0, \Sigma_p)$. We use each template as the mean of an individual-variation Gaussian distribution, $p_i(\mathbf{x}) = N(\mu_i, \Sigma_i)$, to generate the probe set.

The features of the two Gaussian distributions are independent and all dimensions have the same variance, $\sigma_p^2 = 100$ and $\sigma_i^2 = 1$. We show results for dimensional feature-spaces 1 to 3. We compare the predicted and directly calculated CMC curves (mean value of 30 simulations) for 50 and 2000 individuals. Tables 1 and 2 show results for $P(R \leq n)$ where $n = 1, 2, 3, 4$. For Table 2, the predicted CMC values were computed using only a 100 of the 2000 individuals⁴. This was done to show that a small number of individuals (100 for this case) can be used to predict the CMC results for a larger number of individuals.

Table 1. Predicted (PRE) and directly calculated (CAL) cumulative match characteristic values for 50 subjects, for synthetic data.

d	$P(R = 1)$		$P(R \leq 2)$		$P(R \leq 3)$		$P(R \leq 4)$	
	CAL	PRE	CAL	PRE	CAL	PRE	CAL	PRE
1	29.9%	30.4%	50.8%	51.1%	65.7%	66.1%	76.7%	76.9%
2	81.5%	81.3%	95.8%	95.7%	99.0%	98.9%	99.7%	99.7%
3	97.6%	97.2%	99.9%	99.9%	100%	100%	100%	100%

⁴ The 100 individuals are used to estimate Σ_p and Σ_i

Table 2. Predicted (PRE) and directly calculated (CAL) cumulative match characteristic values for 2000 subjects, for synthetic data.

d	$P(R = 1)$		$P(R \leq 2)$		$P(R \leq 3)$		$P(R \leq 4)$	
	CAL	PRE	CAL	PRE	CAL	PRE	CAL	PRE
1	1.4%	1.8%	2.5%	2.9%	3.7%	4.0%	4.8%	5.0%
2	15.7%	15.4%	26.3%	25.9%	34.7%	34.1%	41.3%	40.9%
3	57.7%	56.2%	76.8%	75.2%	85.9%	84.5%	90.9%	89.7%

4.2 Gait Data

Our gait data set is comprised of a set of static body parameters designed for gait recognition from our previous work [1]. While we measure these parameters from video sequences, we also recovered a baseline set of static body parameters from motion-capture data, which we will examine for this paper.

The static body parameters recovered from our motion-capture database, of walking subjects, are four distances: the distance between the head and foot (d_1), the distance between the head and pelvis (d_2), the distance between the foot and pelvis (d_3), and the distance between the left foot and right foot (d_4). These distances are only measured at the maximal separation point of the feet during the double support phase of the gait cycle. Twenty subjects were captured six times yielding a total of 120 instances of each static body parameter. We concatenate the four static body parameters to form a walk vector, $\mathbf{w} = \langle d_1, d_2, d_3, d_4 \rangle$.

In order to compute the directly calculated and predicted CMC curves, the mean of the six data-points from each subject is used as the gallery of templates. The original six data points are used to create six probe set. Each probe set contains one data point per subject. We note that the Σ_i of our gait data did not follow our spherical requirement, so we used the method presented in Section 3 to normalize the data set to satisfy the spherical requirement. Also, the predicted values were calculated using the simplified model of the CMC curve (see Appendix) to show the ability of the simplified model to predict CMC behavior.

Table 3 shows the results of the directly calculated CMC curve and the predicted CMC curve (for $P(R \leq n)$ where $n = 1, 2, 3, 4$). The value of the directly calculated CMC curve is the mean value of the six probe sets with the standard deviation from the mean. The results show that our predicted CMC values are within the standard deviation of the directly calculated CMC curve. In addition, Table 4 is another example of using our CMC model to predict CMC results for subject that are not contain in a database. The table shows predicted values of a CMC curve if 100 subjects were in our gait database using the original 20 subjects to make the calculation.

Tables 1, 2, and 3 show that our model of the CMC curve closely predicts the actual values of the directly calculated curves. From the synthetic data we see that our model is able to predict results at 50 subject and at 2000 subjects. Our model of the CMC curve only depends on knowing the feature-space’s dimension, variance of the population, and variance of the individuals. Which means, if we

Table 3. Predicted (PRE) and directly calculated (CAL) cumulative match characteristic values for 20 subjects, for our gait-recognition data.

	$P(R = 1)$		$P(R \leq 2)$		$P(R \leq 3)$		$P(R \leq 4)$	
	CAL	PRE	CAL	PRE	CAL	PRE	CAL	PRE
w	94.2% \pm 4%	92.1%	99.2% \pm 2%	99.1%	100% \pm 0%	100%	100% \pm 0%	100%

Table 4. Predicted (PRE) cumulative match characteristic values for 100 subjects, for our gait-recognition data.

	$P(R = 1)$	$P(R \leq 2)$	$P(R \leq 3)$	$P(R \leq 4)$
w	71.8%	89.8%	95.7%	98.0%

are able to accurately estimate the population and individual density from a small set of subjects, then we can predict the CMC curve for a larger set of subject that are not present in our data set, as demonstrated in Tables 2 and 4.

5 Conclusion

In this paper we have presented a mathematical formulation of the cumulative match characteristic (CMC) curve given a feature space of biometric measurements. Under our assumptions, we have shown that this mathematical model has the ability to mirror a directly calculated CMC curve for a given number of subjects, and most importantly, it is able to predict CMC curves for a greater number of individuals than which a database actually has. This ability was demonstrated by synthetically generated data and actual gait data from our gait-recognition method.

6 Future Work

As mentioned earlier in the paper, this work is predicated on using the feature space of the biometric technique and the L2norm or Mahalanobis distance to compute similarity within the feature space to model the CMC curve. However, there are a wide range of biometrics techniques that do not have an actual feature space, and similarity is computed with a variety of techniques. In our future work we address creating a CMC model using only the similarity scores. This will make the CMC model more general, and remove the dependence of trying to accurately model the feature-space.

Acknowledgments

Funding for this research was supported in part by the DARPA HID Program, contract #F49620-00-1-0376.

References

1. A. F. Bobick and A. Y. Johnson. Gait recognition using static, activity-specific parameters. *In IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001.
2. R. L. Burden and J. D. Faires. *Numerical Analysis*. PWS-KENT Publishing Company, Boston, 5 edition, 1993.
3. J. P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.
4. A. Jain, R. Bolle, and S. Pankanti. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, Boston, 1999.
5. W. Mendenhall and R. J. Beaver. *Introduction to Probability and Statistics*. Duxbury Press, Belmont, California, 9 edition, 1994.
6. H. Moon and P. J. Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *In Perception*, 30(3):303–321, 2001.

Appendix: Simplified CMC Model

In this appendix, we simplify Equation 5 by removing the multiple integral term. This is accomplished by approximating the probability A^* , in Equation 3, to be the *expected value* of the probability that the measurement of a target’s template is within some distance k of *any* imposter’s template \mathbf{q}_n (see Equation 2), so A^* becomes

$$A^* \approx A(k) = \int_{-\infty}^{+\infty} p_p(\mathbf{q}_n) A(\mathbf{q}_n, k) dq, \quad (7)$$

where

$$A(k) = \frac{V_d \cdot k^d}{2^{d/2} (2\pi)^{d/2} |\Sigma_p|^{1/2}}. \quad (8)$$

Equation 3 now becomes

$$B(R, k) = C_{R-1}^{N-1} A(k)^{R-1} (1 - A(k))^{N-R}. \quad (9)$$

If we integrate over all possible distances k , which a target’s measurement may be at from its template, the probability that the target’s template is exactly rank n is

$$CMC^{**}(R = n) = \int_0^{\infty} B(R, k) p_s(k) dk, \quad (10)$$

which is an approximation of Equation 5. Finally, the probability that a probe data-point’s corresponding gallery data-point is within rank n is

$$CMC(R) = P(R \leq n) = \sum_{i=1}^n CMC^{**}(R = i). \quad (11)$$

Because Equation 11 was created by approximating the probability A^* , it will tend to slightly under estimate the CMC curve at lower ranks and slightly over estimate the CMC curve at higher ranks.