Pre-print of article that appears in T-PAMI, Vol. 33 No. 8 2011.

The published article can be accessed from:
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5740917

Meta-Recognition: The Theory and Practice of Recognition Score Analysis

Walter J. Scheirer, *Member, IEEE,*
Anderson Rocha, *Member, IEEE,*
Ross J. Micheals, *Member, IEEE,*
and Terrance E. Boult, *Member, IEEE*

***Abstract*—In this paper, we define *meta-recognition*, a performance prediction method for recognition algorithms, and examine the theoretical basis for its post-recognition score analysis form through the use of the statistical extreme value theory (EVT). The ability to predict the performance of a recognition system based on its outputs for each match instance is desirable for a number of important reasons, including automatic threshold selection for determining matches and non-matches, and automatic algorithm selection or weighting for multi-algorithm fusion. The emerging body of literature on post-recognition score analysis has been largely constrained to biometrics, where the analysis has been shown to successfully complement or replace image quality metrics as a predictor. We develop a new statistical predictor based upon the Weibull distribution, which produces accurate results on a per instance recognition basis across different recognition problems. Experimental results are provided for two different face recognition algorithms, a fingerprint recognition algorithm, a SIFT-based object recognition system, and a content-based image retrieval system.**

***Index Terms*—Meta-Recognition, Performance Modeling, Multi-Algorithm Fusion, Object Recognition, Face Recognition, Fingerprint Recognition, Content-Based Image Retrieval, Similarity Scores, Extreme Value Theory**

I. Introduction

Recognition in computer vision is commonly defined as submitting an unknown object to an algorithm, which will compare the object to a known set of classes, thus producing a similarity measure to each. For any recognition system, maximizing the performance of recognition is a primary goal. In the case of general object recognition, we do not want an object of a class unknown to the system to be recognized as being part of a known class, nor do we want an object that should be recognized by the system to be rejected as being unknown. In the case of biometric recognition, the stakes are sometimes higher: we never want a misidentification in the case of a watch-list security or surveillance application. With these scenarios in mind, the ability to *predict* the performance of a recognition system on a per instance match basis is desirable for a number of important reasons, including automatic threshold selection for determining matches and non-matches, automatic algorithm selection for multi-algorithm fusion, and further data acquisition signaling — all ways we can improve the basic recognition accuracy.

Meta-recognition is inspired by the multidisciplinary field of meta-cognition study. In the most basic sense, meta-cognition is "knowing about knowing" [1] . For decades, psychologists and cognitive scientists have explored the notion that the human mind has knowledge of its own cognitive processes, and can use it to develop strategies to improve cognitive performance. For example, if a student notices that she has more trouble learning history than mathematics, she "knows" something about her learning ability and can take corrective action to improve her academic performance. We adapt a standard articulation of computational meta-cognition [2], to formally define our meta-recognition:

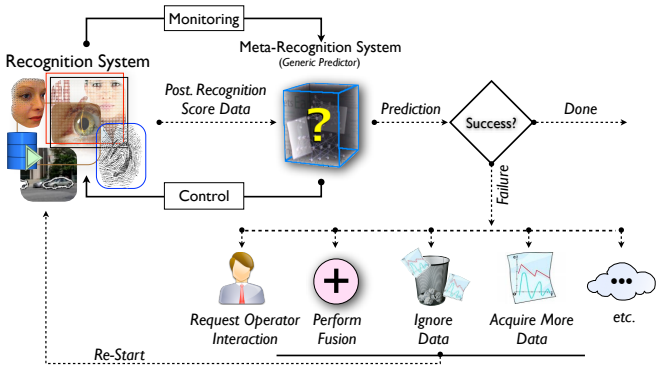


Fig. 1. An overview of the meta-recognition process for post-recognition score analysis. Based upon the scores produced by some recognition system for a single input, a prediction of success or failure is made by the meta-recognition system. Using these predictions, we can take action to improve the overall accuracy of the recognition system.

Definition 1.1 Let X be a recognition system. We define Y to be a *meta-recognition* system when recognition state information flows from X to Y, control information flows from Y to X, and Y analyzes the recognition performance of X, adjusting the control information based upon the observations.

The relationship between X and Y can be seen in Fig. 1, where Y is labeled "Meta-Recognition System". Y can be any approximation of the cognitive process, including a neural network [3], SVM [4], or statistical method. For score-based meta-recognition, the primary approach considered herein, Y observes the recognition scores produced by X, and if necessary, adjusts the recognition decisions and perhaps signals for a specific response action.

Many heuristic approaches could be defined for the meta-recognition process and prior work exists that describes systems that are effectively forms of meta-recognition. Image or sample quality has long stood out as the obvious way of predicting recognition system performance, especially for biometric recognition systems where poor quality images are a frequent occurrence. The National Institute of Standards and Technology (NIST) continues to be the most visible organization promoting quality as a predictor, producing several influential studies [5], [6] that make a strong case for quality as an overall predictor of a system's success. Very bad quality is generally an excellent predictor of failure. However, recent work (also from NIST) suggests that there are cases for challenging the assumption of quality as a universally good predictor - particularly for face recognition.

Walter Scheirer and Terrance Boult are with the Univeristy of Colorado at Colorado Springs and Securics, Inc. Colorado Springs, CO, 80918.
E-mail: lastname@vast.uccs.edu
Anderson Rocha is with the Institute of Computing, University of Campinas (Unicamp), Campinas, Brazil.
Ross Micheals is with the National Institute of Standards and Technology.

Beveridge et al. [7] show that in reasonable systems, different quality assessment algorithms lack correlation in resulting face recognition performance. They also show that images identified as low quality (out of focus) produce better match scores. In [8], Phillips and Beveridge introduce a theory of equivalence in matching and quality, stating that a perfect quality measure for any algorithm would be equivalent to finding a perfect matching algorithm, and thus, bounds are placed on the performance of quality as a predictor. Such a relationship between quality and recognition brings us back to the fundamental issue of matching accuracy. As Beveridge [9] notes, "Quality is not in the eye of the beholder; it is in the *recognition performance figures*!"

Post-recognition score analysis is an emerging paradigm for recognition system prediction, and hence a form of meta-recognition. Fig. 1 depicts the general process, with the analysis occurring after the system has produced a series of distance or similarity scores for a particular match instance. These scores are used as input into a predictor, which will produce a decision of recognition success or failure. This post-recognition classifier can use a variety of different techniques to make its prediction, including distributional modeling and machine learning. Based on the decision of the classifier and not on the original recognition result, action can be taken to lift the accuracy of the system, including enhanced fusion, further data acquisition, or operator intervention. In some cases, the system will be run again to attain a successful recognition result. In the literature, several effective score analysis methods for various matching problems can be found.

Cohort analysis [10], [11], [12], [13], [14], [15] is a *post-verification* (one vs. one matching, as opposed to recognition's one vs. many matching) approach to comparing a claimed object against its neighbors, with many *ad hoc* variations on how to use that cohort information for weighting the results. Some cohort approaches for verification consider scaling by verification scores in a likelihood ratio-like test [10], [12], [13]. More recent work for multibiometric fusion for verification [11], [14], [15] models a cohort class as a distribution of scores from a pre-defined "cohort gallery" and then uses this information to normalize the data. This allows for an estimate of valid "score neighbors", with the expectation that on any match attempt, a claimed object will be accompanied by its cohorts in the sorted score list with a high degree of probability.

While cohort research exists for verification, it is possible to apply a normalization-based cohort methodology to recognition. However, recognition cannot have a consistent pre-defined cohort to compare against during matching. Rather different dynamically varying "cohorts" would likely result for the same individual. One adaptation, used by [14], [15] (and used as a baseline method in this paper) is to treat the entire enrollment gallery as the cohort, leading those authors to observe: "When the cohort models used are the models in the gallery (also known as enrollee or client models) other than the claimed model, one effectively performs identification in the verification mode." While effective and intuitive, normalization-based cohort analysis has lacked a theoretical basis.

Extreme Value Theory as a predictor for vision applications has appeared before, but not for the typical articulation of the recognition problem. For biometric verification, Shi et al. [16] choose to model genuine and impostor distributions using the General Pareto Distribution (GPD). This work makes the important observation that the tails of each score distribution contain the most relevant data to defining each distribution considered for prediction (and the associated decision boundaries), which are often difficult to model — thus the motivation for using EVT. For hyperspectral and radar target detection, GPD has also been applied to isolate extrema within a potential target sample [17]. That work attempts to develop an automatic thresholding scheme, which is an immediate application of any score based prediction system.

First introduced by Li et al. [18], and subsequently used for a variety of biometric prediction applications in [3], [4], [19], machine learning-based post-recognition score analysis has been shown to be very effective. In essence, this technique "learns" from the tails of score distributions in order to construct a classifier that can return a decision of recognition failure or recognition success. Classifiers have been constructed using a variety of features computed from the scores produced by a recognition system. These techniques show much promise for predicting recognition system performance, and for improving [19] recognition results, but have lacked a theoretical foundation.

Thus far, a theoretical explanation of why post-recognition score analysis (including cohort analysis) is effective for per instance matching has yet to be presented. In this paper, we develop a statistical theory of post-recognition score analysis derived from the extreme value theory. This theory generalizes to all recognition systems producing distance or similarity scores over a gallery of known images. Since the literature lacks a specific term for this sort of prediction, we term this work *meta-recognition*. In conjunction with the theory of meta-recognition for post-recognition score analysis, we go on to develop a new statistical classifier based upon the Weibull distribution that produces accurate results on a per instance recognition basis. Experimental results are presented for two different face recognition algorithms, a fingerprint recognition algorithm, a SIFT-based object recognition system, and a content-based image retrieval system.

We organize the rest of this paper as follows. In Section II, we discuss the use of statistical modeling approaches for meta-recognition and also introduce a classification technique for meta-recognition using statistical extreme value theory. In Section III we present experimental results for our statistical predictor on a variety of score data. In Section IV, we draw some conclusions and discuss future directions.

II. Meta-Recognition via Extreme Value Theory

A. Recognition Systems

There are multiple formal ways to define what exactly a "recognition" task is. In [16], Shi et al. define biometric recognition as a hypothesis testing process. In [20], Lowe describes object recognition as a feature vector comparison process requiring a large database of known features and

a distance metric. For this work, we consider the general definition of Shakhnarovich et al. [21], where the task of a recognition system is to find the class label c^*, where p_k is an underlying probability rule and p_0 is the input distribution, satisfying

$$c^* = \underset{class\ c}{\operatorname{argmax}}\, Pr(p_0 = p_c) \tag{1}$$

subject to $Pr(p_0 = p_c^*) \geq 1 - \delta$ for a given confidence threshold δ, or to conclude the lack of such a class (to reject the input). We define $probe$ as the input image p_0 submitted to the system with its corresponding class label c^*. Similarly, we define $gallery$ to be all the classes c^* known to the recognition system. We call this rank-1 recognition because if we sort the class probabilities, the recognition is based on the highest value. One can generalize the concept of recognition, as is common in content-based image retrieval and some biometrics problems, by relaxing the requirement for success to having the correct answer in the top K responses. For analysis, presuming the ground-truth is known, one can define the overall match and non-match distributions for recognition and the per-instance post-recognition distributions (see Fig. 2).

Many systems replace the probability in the above definition with a more generic "score", for which argmax produces the same answer when the posterior class probability is monotonic with the score function. For an operational system, a threshold t_0 on the similarity score s is set to define the boundary between proposed matches and proposed non-matches. The choice of t_0 is often made empirically, based on observed system performance. Where t_0 falls on each tail of each overall distribution establishes where *False Rejection* (Type I error: the probe has a corresponding entry in the gallery, but is rejected) or *False Recognition* (Type II error: the probe does not have a corresponding entry in the gallery, but is incorrectly associated with a gallery entry) will occur. The post-recognition scores in Fig. 2 yield a False Rejection for the t_0 shown. In general, setting a fixed threshold, t_0, on similarity scores produces a recognition confidence δ that varies with each probe.

Based on these definitions, the questions for meta-recognition are: *Can we recognize, in some automated fashion, if a recognition system result is a success or a failure? If so, can we quantify the probability of success or failure?*

B. The Theoretical Basis of Meta-Recognition

As defined in Section II-A, one can map almost any recognition task into the problem of determining "match" scores between the input data and some class descriptor, and then determining the most likely class. Success in a recognition system occurs when the match is the top score. Failure in a recognition system occurs when the match score is not the top score (or not in the top K, for more general rank-K recognition). This must be done for a single probe, and not the overall "match/non-match" distributions, such as those in [16] and [22], which combine scores and performance over many probes. Rather, meta-recognition is done using a single probe, which means it is producing at most one match score mixed in with a larger set of non-match scores.

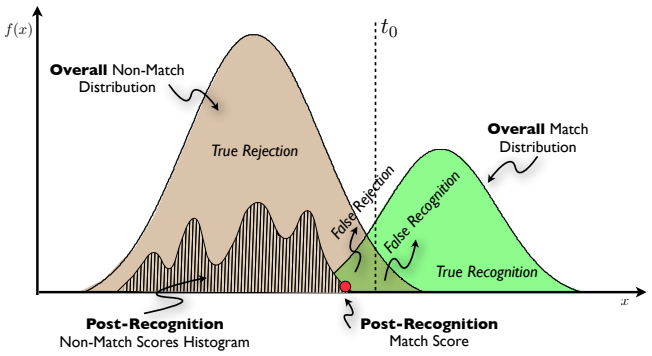


Fig. 2. The match and non-match distributions for the recognition problem. A threshold t_0 applied to the score determines the decision for recognition or rejection. Where the tails of the two distributions overlap is where we find *False Rejections* and *False Recognition*. Embedded within the overall distribution is shown a particular set of post-recognition scores, with one match (falsely rejected by the threshold t_0) and many non-match samples.

Because each recognition instance produces many non-match scores, we can formalize our meta-recognition problem as determining if the top K scores contain an outlier with respect to the current probe's non-match distribution. In particular, let $\mathcal{F}(p)$ be the distribution of the non-match scores that are generated by the matching probe p, and $m(p)$ to be the match score for that probe. In addition, let $S(K) = s_1 \ldots s_K$ be the top K sorted scores. We can formalize the null hypothesis H_0 of our prediction for rank-K recognition as:

$$H_0(failure) : \forall x \in S(K), x \in \mathcal{F}(p), \tag{2}$$

If we can reject H_0 (failure), then we predict success.

While previous researchers have formulated recognition as hypothesis testing given the individual class distributions [21], that approach presumes good models of distributions for each match/class. For a single probe we cannot effectively model the "match" distribution as we only have *one* sample per probe. Assuming a consistent distribution across all probes is dubious.

This is a key insight: *we don't have enough data to model the match distribution, but we have n samples of the non-match distribution — generally enough for good non-match modeling and outlier detection. If the best score is a match, then it should be an outlier with respect to the non-match model.*

As we seek a more formal approach, the critical question then becomes how to model $\mathcal{F}(p)$, and what hypothesis test to use for the outlier detection. Various researchers have investigated modeling the overall non-match distribution [22], developing a binomial model. Our goal, however, is not to model the whole non-match distribution over the entire population, but rather to model the tail of what exists for a single probe comparison. The binomial models developed by [22] account for the bulk of the data, but have problems in the tails. They are not a good model for a particular probe.

An important observation here is that the non-match distribution we seek to model is actually a sampling of scores, one or more per "class," each of which is itself a distribution of potential scores for this probe versus the particular class. Since we consider the upper tail, the top n scores, there is a

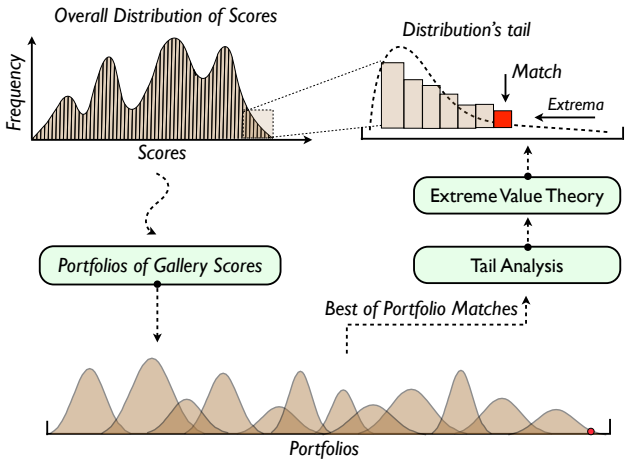


Fig. 3. Why meta-recognition is an extreme value problem. Consider a collection of portfolios composed of subsets of the gallery, each of which produces scores. One portfolio contains a match-score (red), the rest are non-matching scores (brown). The best of the best of the portfolio scores are those that show up in the tail of the post-recognition score distribution — leaving us with an extreme value problem. The best score in the tail is, if a match, an outlier with respect to the EVT model of the non-match data.

strong bias in the samplings that impact the tail modeling; we are interested only in the top scores.

Extreme value problems consider extreme deviations from the median of probability distributions. Thus, it appears intuitive to claim that any analysis considering the tail of a distribution is an extreme value problem. Recent work [17] looking at target detection score spaces relies on this intuition, but does not formally explain why extreme value theory applies to the tails of those score distributions. Just being in the tail is not sufficient to make this an extreme value problem, as one can consider the top N samples from any particular distribution D, which by definition fit distribution D and not any other distribution. Subsequently, the consideration of tail data is not sufficient justification to invoke the extreme value theorem.

The Extreme Value Theorem, also known as the Fisher-Tippet Theorem[23] states:

Extreme Value Theorem 2.1 Let $(s_1, s_2, \ldots)$ be a sequence of i.i.d samples. Let $M_n = \max\{s_1, \ldots, s_n\}$. If a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and

$$\lim_{x\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x) \tag{3}$$

then if F is a non-degenerate distribution function, it belongs to one of three extreme value distributions.

To see that recognition is an extreme value problem in a formal sense, we can consider the recognition problem as logically starting with a collection of *portfolios* (here we borrow the term from financial analysis, where EVT is broadly applied). Each portfolio is an independent subset of the gallery or recognition classes. This is shown in Figure 3. From each portfolio, we can compute the "best" matching score in that portfolio. We can then collect a subset of *all* the scores that are maxima (extrema) within their respective portfolios. The tail of the post-match distribution of scores will be the best scores from the best of the portfolios. Looking at it this way we have shown that modeling the non-match data in the tail is indeed an extreme value problem.

Thus, a particular portfolio is represented as the sampling $(s_1, s_2, \ldots)$ drawn from an overall distribution of scores S. The maximum of a portfolio is a single sample from the distribution function $F(x)$. Theorem 2.1 tells us that a large set of individual maxima M_n from the portfolios must converge to an extreme value distribution. As portfolio maxima fall into the tail of S, they can be most accurately modeled by the appropriate extreme value distribution. The assumptions necessary to apply this for a recognition problem are that we have sufficiently many classes for the portfolio model to be good enough for the approximation in the limit to apply, and that the portfolio samples are *i.i.d.* (relaxed below).

The EVT is analogous to a central limit theorem, but tells us what the distribution of extreme values should look like as we approach the limit. Extreme value distributions are the limiting distributions that occur for the maximum (*or* minimum, depending on the data representation) of a large collection of random observations from an arbitrary distribution. Gumbel [24] showed that for any continuous and invertible initial distribution, only three models are needed, depending on whether the maximum or the minimum is of interest, and also if the observations are bounded from above or below. Gumbel also proved that if a system or part has multiple failure modes, the failure is best modeled by the Weibull distribution. The resulting three types of extreme value distributions can be unified into a generalized extreme value (GEV) distribution given by

$$GEV(t) = \begin{cases} \frac{1}{\lambda} e^{-v^{-1/k}} v^{-(1/k+1)} & k \neq 0 \\ \frac{1}{\lambda} e^{-\left(x + e^{-x}\right)} & k = 0 \end{cases} \tag{4}$$

where $x = \frac{t-\tau}{\lambda}, v = (1 + k\frac{t-\tau}{\lambda})$ where k, λ, and τ are the shape, scale, and location parameters respectively. Different values of the shape parameter yield the extreme value type I, II, and III distributions. Specifically, the three cases $k = 0, k > 0$, and $k < 0$ correspond to the Gumbel (I), Frechet (II), and Reversed Weibull (III) distributions. Gumbel and Frechet are for unbounded distributions and Weibull for bounded.

If we presume that match scores are bounded, then the distribution of the minimum (or maximum) reduces to a Weibull (or Reversed Weibull) [25], independent of the choice of model for the individual non-match distribution. For most recognition systems, the distance or similarity scores are bounded from both above and below. If the values are unbounded, the GEV distribution can be used. Most importantly, we don't have to assume distributional models for the match or non-match distributions. Rephrasing, no matter what model best fits each non-match distribution, be it a truncated binomial, a truncated mixture of Gaussians, or even a complicated but bounded multi-modal distribution, with enough samples and enough classes *the sampling of the top-n scores always results in a EVT distribution, and is Weibull if the data are bounded.*

Given the potential variations that can occur in the class for which the probe image belongs, there is a distribution of scores that can occur for each of the classes in the gallery. Figure 3 depicts the recognition of a given probe image as implicitly sampling from these distributions. Our method takes the tail of these scores, which are likely to have been sampled from the extrema of their underlying portfolios, and fits a Weibull distribution to that data. Given the Weibull fit to the data, we can answer the meta-recognition question using a hypothesis test to determine if the top score is an outlier by considering the amount of the cumulative distribution function (CDF) that is to the right of the top score, or determine the probability of failure directly from the inverse CDF of that score.

While the classic EVT is presented assuming i.i.d. samples, it can be generalized to the weaker assumption of exchangeable random variables [26], resulting in at most a mixture of underlying EVT distributions. Consider the special case of identically distributed (but not independent) exchangeable variables drawn from the same EVT family, possibly with different parameters. With a mild assumption of bounded mean-square convergence, the underlying distribution even under exchangeable random variables is the same distribution as the classic case (see Theorems 2.1, 2.2 and Corollary 2.2 of [26]). For the recognition problem, it is quite reasonable to presume that the scores generated from matching one class versus another generates a distribution with a *form* that does not depend on the classes involved, even if the parameters do. This is a rather weak assumption. The distribution can be any form and each pair of classes can have any set of parameters, as long as the sampling is exchangeable (for example, later samples do not depend on values from earlier samples). We don't need to know the form or the parameters, we just must assume it exists and is a proper distribution.

C. Weibull-based Statistical Meta-Recognition

As we propose to use the consistency of the Weibull model of the non-match data to the top scores, an issue that must be addressed in statistical meta-recognition is the impact of any outliers on the fitting. For rank-1 fitting, this bias is easily reduced by excluding the top score and fitting to the remaining $n-1$ scores from the top n. If the top score is an outlier (recognition worked), then it does not impact the fitting. If the top score was not a match, including the recognition in the fitting will not only bias the distribution to be broader than it should, but will also increase the chance that the system will classify the top score as a failure. For rank-K recognition, we employ a cross-validation approach for the top-K elements, but for simplicity herein we focus on the rank-1 process. We must also address the choice of n, the tail size to be used.

Given the above discussion we can implement rank-1 meta-recognition as shown in Algorithm 1. An inverse Weibull distribution allows for the estimation of the "confidence" likelihood of a particular measurement being drawn from a given Weibull distribution, which is how we will test for "outliers". In this formulation, δ is the recognition confidence or hypothesis test "significance" level threshold. While we will show full curves in the experiments (Section III), good performance is often achieved using $\delta = 1 - 10^{-8}$.

Algorithm 1 Rank-1 Statistical Meta-Recognition.

Require: A collection of similarity scores S

1: **Sort** and retain the n largest scores, $s_1, \ldots, s_n \in S$;
2: **Fit** a GEV or Weibull distribution W to $s_2, \ldots, s_n$, skipping the hypothesized outlier;
3: **if** $Inv(W(s_1)) > \delta$ **then**
4: s_1 is an outlier and we reject the failure prediction (null) hypothesis H_0.
5: **end if**

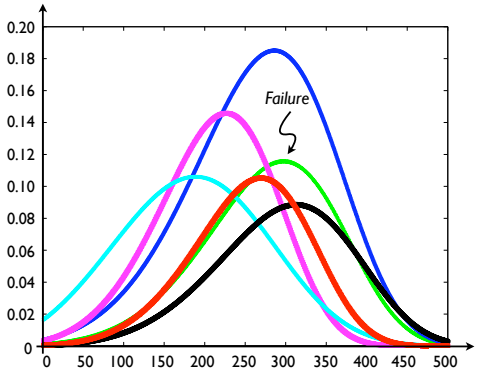


Fig. 4. Weibull distributions recovered from six different real-matches (from the finger LI set of the NIST BSSR1 multibiometric data set), one is a failure (not rank-1 recognition), five are successes. Per-instance success and failure distributions are not distinguishable by shape or position. In this example, the green distribution is a recognition failure, while the rest are successes.

It is desirable that the meta-recognition methodology does not make any assumptions about the arithmetic difference between low matching and high non-matching scores. If the data satisfied the assumption of high arithmetic difference among the match and non-match scores, a simple threshold would suffice for meta-recognition. As a matter of fact, our meta-recognition approach shows good performance in many different scenarios — even with scores that are almost tied. Fig. 4 depicts six different Weibull distributions recovered from real matching instances of the fingerprint LI subset of NIST's BSSR1 [27] multibiometric data set. Visually, it is unclear which Weibull distributions are correct matches, and which are not. It is not the mean or the shape, but the outlier test that allows our Weibull-based meta-recognition approach to make the distinction.

III. Meta-Recognition: Experiments & Validation

A. Meta-Recognition Error Trade-off Curves

In order to assess the performance of the prediction approach we introduce in this paper, we require an analysis tool similar to a detection error trade-off curve, which allows us to vary parameters to gain a broad overview of the system behavior. We can calculate a "Meta-Recognition Error Trade-off Curve" (MRET) from the following four cases:

C_1 "**False Accept**", when meta-recognition predicts that the recognition system will succeed but the rank-1 score is not correct.

C_2 "**False Reject**", when meta-recognition predicts that the recognition system will fail but rank-1 is correct.

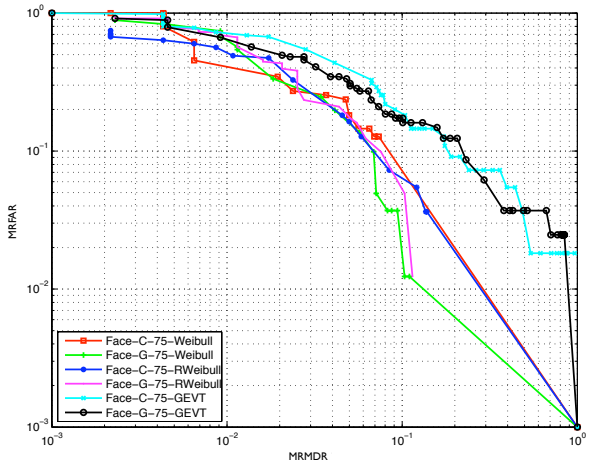


Fig. 5. MRET curves for comparing GEVT, reversed Weibull and Weibull-based predictions using the BSSR1 data set algorithms face C and face G. Weibull clearly outperforms the more general GEVT. Weibull and reversed Weibull are close. The tail size of 75 used for Weibull fitting is 14.5% of the total scores.

C_3 "**True Accept**", when both the recognition system and meta-recognition indicate a successful match.

C_4 "**True Reject**", when meta-recognition predicts correctly that the underlying recognition system is failing.

We calculate the Meta-Recognition False Accept Rate (MRFAR), the rate at which meta-recognition incorrectly predicts success, and the Meta-Recognition Miss Detection Rate (MRMDR), the rate at which the meta-recognition incorrectly predicts failure, as

$$MRFAR = \frac{|C_1|}{|C_1| + |C_4|}, \quad MRMDR = \frac{|C_2|}{|C_2| + |C_3|}. \quad (5)$$

This representation is a convenient indication of meta-recognition performance, and we use it to express all the results we present in this paper. The MRFAR and MRMDR can be adjusted via thresholding applied to the predictions to build the curve. Just as one uses a traditional DET or ROC curve to set verification system parameters, the meta-recognition parameters can be tuned using the MRET.

B. Statistical Meta-Recognition Results

In practice, statistical meta-recogntion is an excellent predictor of recognition algorithm success or failure. Table I lists the complete breakdown for the experiments presented in this section. Each experiment is associated with scores from a particular recognition algorithm run on a standard data set. We consider all positive and negative match instances available in our data as individual tests, with MRET curves generated by considering all of the individual meta-recognition results for a particular algorithm and data set. Note the wide variation in total tests (500 - 1624). This affects the shape of the curves in Figs. 5 - 7, with more data producing a smoother curve.

Here we draw a number of interesting conclusions from a variety of meta-recognition experiments. First, we confirm our hypothesis that the Weibull distribution is the most suitable distribution for statistical meta-recognition. The theory of

Data	Rank-1 Correct	Rank-1 Incorrect	Total Tests
BSSR1 C Multibiometric	462	55	517
BSSR1 G Multibiometric	436	81	517
BSSR1 LI Multibiometric	448	69	517
BSSR1 RI Multibiometric	481	36	517
FERET EBGM	935	269	1204
ALOI Illum. SIFT	227	273	500
"Corel Relevants" bic	1360	264	1624
"Corel Relevants" ccv	1189	435	1624
"Corel Relevants" gch	1163	461	1624
"Corel Relevants" lch	1116	508	1624

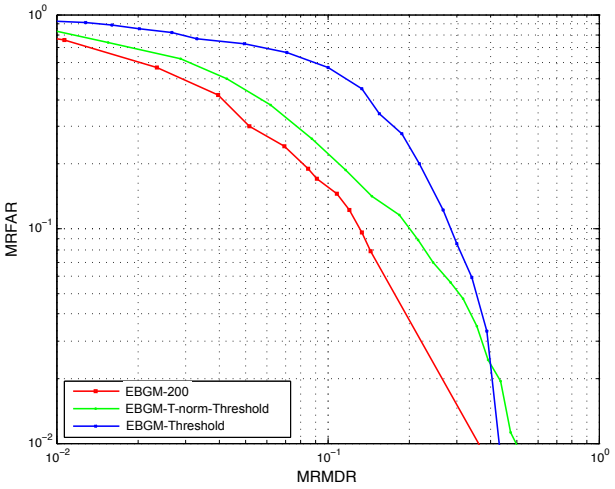
TABLE I
DATA BREAKDOWN FOR THE META-RECOGNITION EXPERIMENTS.

Section II-B requires a statistical significance of deviation from the model for classification. Section II-C defined a formal statistical test for such significance. To analyze the choice of model, including Weibull, Reversed Weibull, and GEVT, we used the face-recognition algorithms from the NIST BSSR1 multibiometric score set; we show the comparison in Fig. 5. To interpret this plot (and the following MRET curves), it must be understood that points approaching the lower left corner minimize both the MRFAR and MRMDR errors. In Fig. 5, the two Weibull and two Reversed Weibull curves reflect higher accuracy, when compared to the two GEVT curves. This is consistent with our earlier claim in Section II-B about our choice of distribution. Because most recognition scores are bounded from both above and below, Weibull is the most appropriate EVT distribution for modeling the recognition problem and is empirically more accurate than the GEVT.

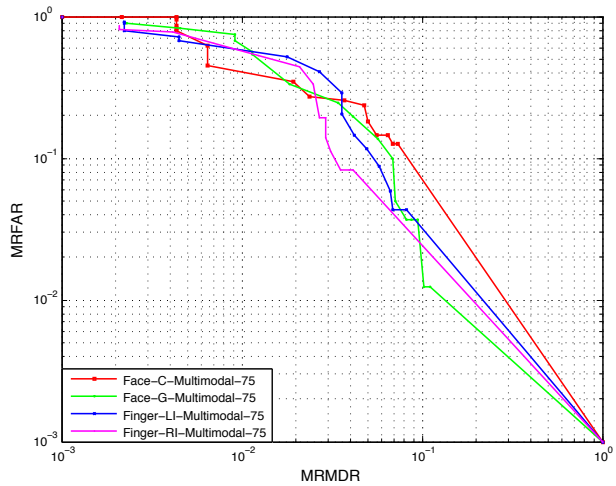
Second, we confirm that statistical meta-recognition is significantly better than a standard threshold test over the original score data and T-norm scores [14] [15]. Along with the meta-recognition results for the Elastic Bunch Graph Matching (EBGM) [28] algorithm from the CSU Facial Identification Evaluation System [29], the data for a trivial form of prediction is also depicted in Fig. 6(a) (labeled "Threshold"). The comparison curve is generated by varying a series of thresholds (from 0 to 0.99, at intervals of 0.01), with each score compared against each threshold point. If the original score is greater than the threshold for a particular point, then we consider this a prediction of success, otherwise, we predict failure. We compare this prediction to the ground-truth for every score series, thus building the MRET curve.

T-norm scores were generated, following [14], by considering the hypothesized non-match scores (all scores after the top score) as the data used to calculate the necessary statistics. In a 10-fold cross validation approach, we randomly selected cohorts of size $|\mathcal{F}(p)| - 100$ for each match instance and normalized the entire score series based on the calculated statistics for the cohort. Each normalized score was then scaled to bring it between 0 and 0.99, and the above threshold prediction was applied to generate the MRET curve data. In Fig. 6(a), each point on the T-norm curve represents the mean of all 10 MRFAR and MRMDR values. Error bars were smaller than the plot value symbols and are not shown.

Fig. 6(a) shows that the EVT-based meta-recognition technique (labeled EBGM-200) significantly outperforms the

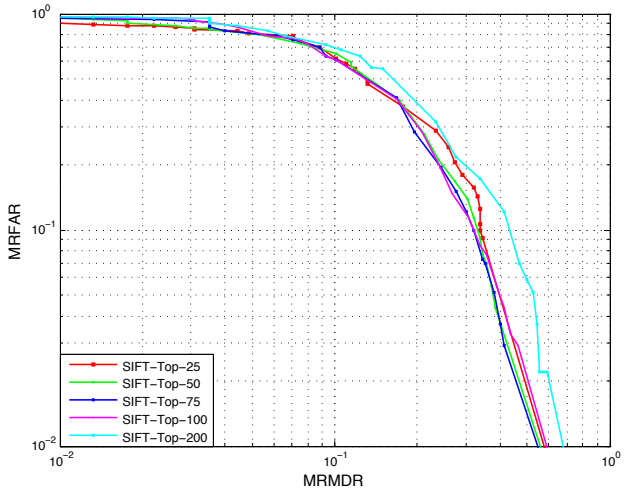


(a) EBGM

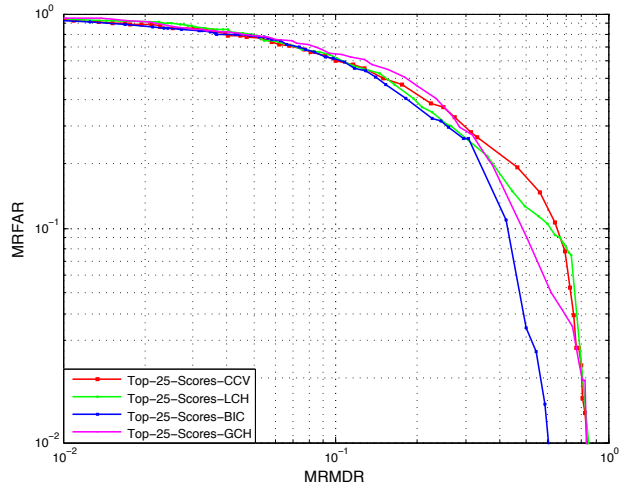


(b) BSSR1 Multibiometric Set

Fig. 6. MRET curves for biometric recognition algorithms. For EBGM (a) best tail size of 200 scores (17% of the total scores) is shown, with comparison curves for basic thresholding over original scores and T-norm scores. The data set is the entire FERET set. The true multibiometric set of BSSR1 (b), maintains gallery consistency across the different algorithms. The tail size of 75 used for Weibull fitting is 14.5% of the total scores.



(a) SIFT



(b) Four CBIR Descriptors

Fig. 7. MRET curves for object recognition algorithms. For SIFT (a), EMD is the distance metric and the data set is the illumination direction subset of ALOI. Tail sizes used for Weibull fitting vary from 25 scores (5% of the total scores) to 200 scores (40% of the total scores). For the CBIR descriptors (b), the data set is "Corel Relevants". The tail size of 25 used for Weibull fitting is 50% of the total scores.

pure threshold technique (labeled EBGM-Threshold) as well as the T-norm based thresholding (labeled EBGM-T-norm-Threshold). The equal error rate (the point at which MRFAR and MRMDR errors are equal) for the EBGM-200 curve is roughly 10%, meaning that just 1 out of 10 meta-recognition instances will incorrectly predict success or failure for this algorithm and tail size. The EBGM-Threshold curve has an equal error rate of 20%, and is much worse at other points along the curve in comparison to the meta-recognition curve. Interestingly, the EBGM-T-norm-Threshold curve shows higher accuracy than the EBGM-Threshold curve, but is still always worse in accuracy compared to the meta-recognition EBGM-200 curve.

Third, we evaluate our only parameter for the statistical meta-recognition process: tail size. In all of the plots, we have used the notation DATA-tailsize to show the tail size used for the Weibull fitting piece of our algorithm. In practice, the selection of the tail size is very important for meta-recognition accuracy. The best performing tail size is found to be a function of the gallery size; as the gallery grows, so too does the amount of tail data we must consider. To emphasize this point, tail statistics are given in the figure captions.

Fourth, we select a series of algorithms and data sets that reflect a variety of typical recognition cases - including those where fusion is applicable. Fig. 6(b) depicts results for the NIST BSSR1 multibiometric score set, including scores from 2 face recognition algorithms and 1 fingerprint recognition algorithm (for two index fingers, labeled LI and RI). In this true multibiometric subset, the gallery is consistent across all algorithms, making it possible to fuse across all of the data to improve recognition results. A score level fusion system can incorporate meta-recognition to identify algorithms that have failed for a particular recognition instance, and remove them for consideration before any fusion takes place.

We are also not just limited to biometric recognition algorithms. Fig. 7(a) depicts results for a SIFT-based approach [20] for object recognition on the illumination direction subset of the Amsterdam Library of Objects (ALOI) set [30], while Fig. 7(b) depicts results for four different Content-Based Image Retrieval approaches [31] on the "Corel Relevants" data set [32]. As in Fig. 6(b), Fig. 7(b) shows good potential for score level fusion between CBIR descriptors. This wide variety of experiments highlights meta-recognition's applicability as a general technique for many different computer vision problems.

IV. Conclusion

In this paper, we have introduced meta-recognition, a performance prediction method for recognition algorithms that allows us to observe the results of the recognition process and, if necessary, adjust the recognition decisions. Using Extreme Value Theory concepts, we have presented a theoretical explanation of why meta-recognition for post-recognition score analysis is effective. We showed that this theory generalizes to all systems that produce distance or similarity scores over a gallery of known examples. The concept of meta-recognition can be applied broadly, and we encourage researchers in general object recognition, AI and other areas looking at recognition to consider it for their domains.

To perform statistical meta-recognition, we have focused on modeling the tail of the non-match distribution of scores. For that, we considered this problem as a collection of portfolios composed of subsets of scores from the overall distribution of scores from the gallery. With this in mind, we have introduced a new statistical classifier that can predict the success or failure of a recognition system's output based on the Weibull distribution. This classifier yields accurate results on a per instance recognition basis without any prior information.

The introduced techniques allow us to make recognition decisions without the need of any *a priori* score threshold selection. For future directions, we intend to explore new applications for the proposed techniques, incorporate meta-recognition into fusion frameworks for recognition systems [33], as well as continue to investigate possible enhancements to improve the accuracy of meta-recognition.

Acknowledgment

Supported in part by ONR STTR N00014-07-M-0421, ONR SBIR N00014-09-M-0448, NSF PFI Award #065025, and FAPESP Award #2010/05647-4. We also thank J. Ross Beveridge, who provided valuable feedback on early drafts of this work.

References

[1] J. Flavell and H. Wellman, "Metamemory," in *Perspectives on the Development of Memory and Cognition*, J. R. V. Kail and J. W. Hagen, Eds. LEA, 1988, pp. 3–33.
[2] M. Cox, "Metacognition in Computation: a Selected Research review," *Artificial Intelligence*, vol. 169, no. 2, pp. 104–141, 2005.
[3] T. Riopka and T. Boult, "Classification Enhancement via Biometric Pattern Perturbation," in *IAPR AVBPA*, vol. 3546, 2005, pp. 850–859.
[4] W. Scheirer, A. Bendale, and T. Boult, "Predicting Biometric Facial Recognition Failure With Similarity Surfaces and Support Vector Machines," in *Proc. of the IEEE Workshop on Biometrics*, 2008.
[5] E. Tabassi, C. Wilson, and C. Watson, "Fingerprint Image Quality, NFIQ," in *Nat. Inst. of Standards and Technology, NISTIR 7151*, 2004.
[6] P. Grother and E. Tabassi, "Performance of Biometric Quality Evaluations," *IEEE TPAMI*, vol. 29, no. 4, pp. 531–543, 2007.
[7] J. R. Beveridge, G. Givens, P. J. Phillips, and B. Draper, "Focus on Quality, Predicting FRVT 2006 Performance," in *Intl. Conf. on Automatic Face and Gesture Recognition*, 2008.
[8] P. Phillips and J. R. Beveridge, "An Introduction to Biometric-completeness: The Equivalence of Matching and Quality," in *IEEE BTAS*, 2009.
[9] J. R. Beveridge, "Face Recognition Vendor Test 2006 Experiment 4 Covariate Study," 2008, presentation at 1st MBGC Kick-off Workshop.
[10] S. Furui, "Recent Advances in Speaker Recognition," *Pat. Rec. Letters*, vol. 18, no. 9, pp. 859 – 872, 1997.
[11] S. Tulyakov, Z. Zhang, and V. Govindaraju, "Comparison of Combination Methods Utilizing t-normalization and Second Best Score Models," in *Proc. of the IEEE Workshop on Biometrics*, 2008.
[12] G. Aggarwal, N. Ratha, R. Bolle, and R. Chellappa, "Multi-biometric Cohort Analysis for Biometric Fusion," in *Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing*, 2008.
[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
[14] N. Poh, A. Merati, and J. Kittler, "Adaptive Client-Impostor Centric Score Normalization: A Case Study in Fingerprint Verication," in *IEEE BTAS*, 2009.
[15] ——, "Making Better Biometric Decisions with Quality and Cohort Information: A Case Study in Fingerprint Verication," in *EUSIPCO*, 2009.
[16] Z. Shi, F. Kiefer, J. Schneider, and V. Govindaraju, "Modeling Biometric Systems Using the General Pareto Distribution (GDP)," in *Proc. of the SPIE*, vol. 6944, 2008, pp. 69 440O–69 440O–11.
[17] J. Broadwater and R. Chellappa, "Adaptive Threshold Estimation Via Extreme Value Theory," *IEEE TSP*, vol. 58, no. 2, 2010.
[18] W. Li, X. Gao, and T. Boult, "Predicting Biometric System Failure," in *IEEE CIHSPS*, 2005.
[19] W. Scheirer and T. Boult, "A Fusion-Based Approach to Enhancing Multi-Modal Biometric Recognition System Failure Prediction and Overall Performance," in *IEEE BTAS*, 2008.
[20] D. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[21] G. Shakhnarovich, J. Fisher, and T. Darrell, "Face Recognition From Long-term Observations," in *ECCV*, 2002, pp. 851–868.
[22] P. Grother and P. Phillips, "Models of Large Population Recognition Performance," in *IEEE CVPR*, 2004, pp. 68–75.
[23] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications*, 1st ed. World Scientific Publishing Co., 2001.
[24] E. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications*, ser. 33. Washington, D.C.: U.S. GPO, 1954, no. National Bureau of Standards Applied Mathematics.
[25] NIST, *NIST/SEMATECH e-Handbook of Statistical Methods*, ser. 33. U.S. GPO, 2008.
[26] S. Berman, "Limiting Distribution of the Maximum Term in Sequences of Dependent Random Variables," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 894–908, 1962.
[27] "NIST Biometric Scores Set," 2004, http://www.itl.nist.gov/iad/894.03/biometricscores/.
[28] K. Okada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test," in *Face Recognition: From Theory to Applications*, H. Wechsler, P. Phillips, V. Bruce, F. F. Soulie, and T. Huang, Eds. Springer-Verlag, 1998, pp. 186–205.
[29] D. Bolme, J. R. Beveridge, M. Teixeira, and B. Draper, "The CSU Face Identification Evaluation System: Its Purpose, Features, and Structure," in *ICVS*, 2003, pp. 304–313.
[30] J. Geusebroek, G. Burghouts, and A. Smeulders, "The Amsterdam Library of Object Images," *IJCV*, vol. 61, no. 1, pp. 103–112, 2005.
[31] J. Almeida, A. Rocha, R. Torres, and S. Goldenstein, "Making Colors Worth More Than a Thousand Words," in *ACM SAC*, 2008, pp. 1179–1185.
[32] R. Stehling, M. Nascimento, and A. Falcão, "A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification," in *ACM CIKM*, 2002, pp. 102–109.
[33] W. Scheirer, A. Rocha, R. Micheals, and T. Boult, "Robust Fusion: Extreme Value Theory for Recognition Score Normalization," in *ECCV*, 2010, pp. 481–495.