Learning for Meta-Recognition

Walter J. Scheirer, *Member, IEEE,* Anderson Rocha, *Member, IEEE,*
Jonathan Parris, *Student Member, IEEE,* Terrance E. Boult, *Member, IEEE*

Abstract—In this work, we consider meta-recognition, an approach for post-recognition score analysis, whereby a prediction of matching accuracy is made from an examination of the tail of the scores produced by a recognition algorithm. This is a general approach that can be applied to any recognition algorithm producing distance or similarity scores. In practice, meta-recognition can be implemented in two different ways: a statistical fitting algorithm based on the Extreme Value Theory, and a machine learning algorithm utilizing features computed from the raw scores. While the statistical algorithm establishes a strong theoretical basis for meta-recognition, the machine learning algorithm is more accurate in its predictions in all of our assessments. In this article, we present a study of the machine learning algorithm and its associated features for the purpose of building a highly accurate meta-recognition system for security and surveillance applications. Through the use of feature- and decision-level fusion, we achieve levels of accuracy well beyond those of the statistical algorithm, as well as the popular "cohort" model for post-recognition score analysis. In addition, we also explore the theoretical question of why machine learning-based algorithms tend to outperform statistical meta-recognition, and provide a partial explanation. We show that our proposed methods are effective for a variety of different recognition applications across security and forensics-oriented computer vision, including biometrics, object recognition, and content-based image retrieval.

Index Terms—Meta-Recognition, Performance Modeling, Object Recognition, Face Recognition, Fingerprint Recognition, Content-Based Image Retrieval, Multi-biometric Fusion, Similarity Scores, Machine Learning

I. INTRODUCTION

THE growing demand for highly accurate surveillance, intelligence, and forensics systems has propelled the unconstrained recognition problem (1:N matching, or *identification*) to the forefront of computer vision research. Over the past decade, excellent progress has been made toward the constrained and unconstrained verification (1:1 matching) problems. For controlled face verification, 99.9% accuracy was achieved for the FRGC set [1]. For controlled fingerprint verification, accuracies between 85.83% and 99.98% have been reported for the FVC2006 set [2]. For uncontrolled face verification, 88.13% accuracy has been reported for the once very difficult LFW set [3]. Verification is a fundamentally easier problem than recognition, as it only considers discrete pairs of samples for matching, with a claimed identity

Author's address: *Dr. Scheirer*, *Mr. Parris* and *Dr. Boult* are with the Department of Computer Science, University of Colorado at Colorado Springs, 1420 Austin Bluffs Parkway, Colorado Springs, CO, USA, 80918.
Dr. Rocha is with the Institute of Computing, University of Campinas, Av. Albert Einstein, 1251 – Campinas, São Paulo, Brazil, 13083-852.
Corresponding authors: Walter J. Scheirer (`wjs3@vast.uccs.edu`) and Anderson Rocha (`anderson@ic.unicamp.br`)
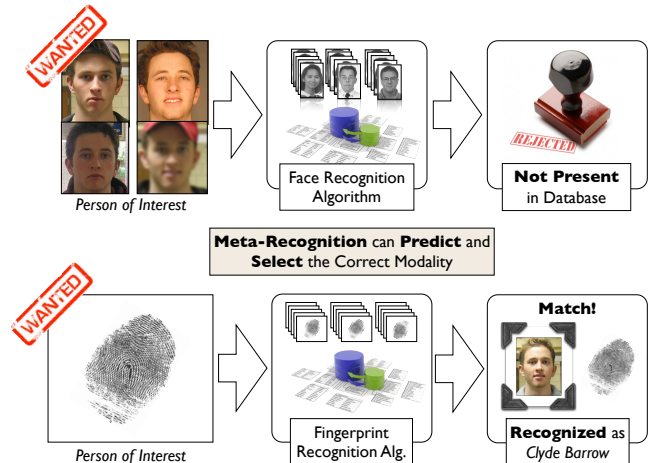
Fig. 1. Our goal in this work is to develop a highly accurate machine learning-based algorithm for *Meta-Recognition*. Meta-Recognition is a post-recognition score analysis technique that predicts when a recognition algorithm is succeeding or failing. In the above example for face recognition, we can see that a number of effects negatively impact recognition accuracy, including blur, illumination, expression, occlusion, shadow, and even changing hairstyles. However, if another more stable modality such as fingerprint is available during the recognition process, a correct recognition result can be achieved. The question then becomes: which modality is correct? The predictions produced by a meta-recognition system can select the correct modality in instances like this.

choosing a comparison class that is known to the matching system. Recognition is made more difficult by the need to identify an unknown class out of the set of known classes. Compounding things further is the overall environment of the unconstrained scenario, where any number of effects (pose, illumination, expression, sensor noise, blur, occlusion, weather, etc.) can impact accuracy. Recognition, in general, is a challenging problem with important consequences for security and forensics applications.

A common approach for improving recognition accuracy is to combine results from a collection of algorithms and/or sensors using score-level fusion [4]. Most of the available fusion works reported in the literature have focused on either combining consistent data to address sensor limitations or limiting the impact of a failing modality when score data is combined. *Meta-Recognition* is a post-recognition score analysis technique that predicts when a recognition algorithm is succeeding or failing. This is very different from any fusion approach that is focused on combining consistent data. If a screening system is being negatively impacted by the environment (Fig. 1), then a biometric (such as multi-view face) that is providing a more consistent answer than another (perhaps fingerprint) does not always mean it should be considered with more emphasis. For instance, if our analysis predicts *success* for one modality and *failure* for the other, we can proceed

with the modality that isn't failing, consistent or not.

A concrete example of this problem comes to us from a 2007 incident in the news [5]. Juan Carlos Ramírez Abadía, a Cali Cartel trafficker, was apprehended under very interesting technological circumstances. Having a penchant for secrecy and disguise, Abadía was well aware of the tools of surveillance deployed by law enforcement, including automated face recognition. He underwent extensive facial surgery and his constantly changing appearance would have easily foiled even state-of-the-art face recognition algorithms, which rely on some consistency between the enrollment images and the input images for matching. Abadía's 2007 arrest in São Paulo, Brazil was made possible after Brazilian authorities provided the American DEA with a voice print for automatic recognition. Voice was a modality that was not "failing," allowing it to indicate Abadía's identity. Another example dates from December 2009, when the Japanese authorities reported the case of a Chinese woman trying to evade the border control by using fake fingerprints [6].

The previous two examples show that plastic surgery and disguise are tools for evading biometric systems. They are part of an ever-growing problem that has attracted the attention of the biometrics community [7], [8]. In the first scenario, if we could *predict* that a face recognition system is failing, and a voiceprint recognition system is succeeding, we could successfully recognize someone like Abadía in an automated fashion – without the face recognition system biasing the final result. Thus, in this article, we turn to post-recognition score analysis to provide such predictions. With meta-recognition, we can produce highly accurate classifiers that can help us solve a case such as Abadía's with no further difficulties[1].

Meta-recognition is formally defined as a control relationship between the post-recognition score analysis and a recognition system [10]:

Definition 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. 2, where Y is labeled "Meta-Recognition System". Y can be one of a number of classification algorithms, such as a neural network [11], support vector machine [12], or the statistical extreme value theory [10]. 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. We note that meta-recognition is different from *meta-analysis* because it does not draw broad conclusions over multiple studies. Instead, it considers recognition on a per instance matching basis.

With several options for actually implementing meta-recognition, the question of which to choose for operational

[1]Or maybe not. Abadía proved to be rather tech savvy in general, having also utilized steganography [9] in his criminal career. It is conceivable that he is a harbinger of clever criminal activity to come.
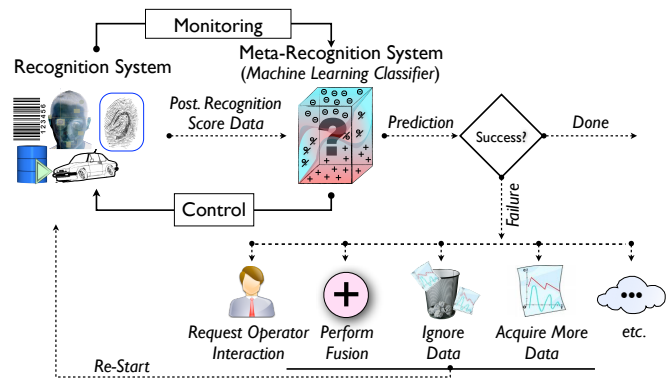

Fig. 2. An overview of the meta-recognition process. Based upon the scores produced by some recognition system for a single input, a prediction of success or failure is made by the meta-recognition system (in this work, we primarily consider machine learning for this). Using these predictions, we can take action to improve the overall accuracy of the recognition system. For instance, if the recognition system has failed to recognize the input image, we can perform better fusion with other collected data by down-weighting or discarding the failing data, ignoring the data, or acquiring more data, giving the recognition system another attempt to recognize the input image.

systems is important. In our prior work, we have explored both pure statistical algorithms and machine learning based algorithms. In [10], we introduced a strong theoretical basis for statistical meta-recognition using the extreme value theory (EVT). The EVT approach produced predictors with accuracies well beyond standard thresholding and cohort thresholding, without the need for training data. However, our machine learning based algorithms [11]–[16] have consistently been more accurate than the EVT approach over numerous experiments. In light of this observation, we have sought a deeper understanding of the underlying feature mechanisms for the machine learning algorithms that lead to higher accuracies.

In this article, we introduce three contributions. First, we provide a study of learning for the purpose of building a highly accurate meta-recognition system that constitutes Y of Def. 1. Second, through the use of feature- and decision-level fusion, we present techniques that achieve levels of accuracy well beyond those of the statistical algorithm, as well as the popular "cohort" model for post-recognition score analysis. Third, we also explore the theoretical question of why the machine learning algorithm tends to outperform the statistical algorithm in many cases of meta-recognition. We show that the introduced methods are effective for a variety of different recognition applications across security and forensics-oriented computer vision.

II. PRIOR WORK

The idea of a post-recognition score analysis to improve the accuracy of pattern recognition is well known. Duda et al. [17] describe classification post-processing to be something similar to our description of meta-recognition in Sec. I:

"A classifier rarely exists in a vacuum. Instead, it is generally to be used to recommend actions (put this fish in this bucket, put that fish in that bucket), each action having an association cost. The post-processor uses the output of the classifier to decide on the recommended action."

Further action, for Duda et al., consists of error rate calculation, risk assessment, context exploitation, fusion of results from multiple classifiers, and complexity assessment. Bishop [18] also hints at the utility of post-recognition score analysis, specifically for error assessment.

Many heuristic approaches could be defined for the post-recognition score analysis 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.

A Study from The National Institute of Standards and Technology (NIST) discusses that very bad quality is generally an excellent predictor of failure [19]. 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. Beveridge et al. [20] show that in reasonable systems, different quality assessment algorithms lack correlation in resulting face recognition performance (indeed images identified as low quality produce better match scores). In [21], 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.

Cohort analysis [22]–[25] is a *post-verification* 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 [22], [23]. More recent work for multi-biometric fusion for verification [24], [25] 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 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 [24] and [25], (and used as a baseline method in this article) 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 lacks a theoretical basis.

Our prior work [10] in meta-recognition analysis developed a strong theoretical model for distance or similarity scores from recognition algorithms based on the statistical extreme value theory. The sampling of the top-n scores (the tail) from a score distribution always results in an EVT distribution and is Weibull if the data are bounded. From this observation, we were able to build a statistical predictor by fitting a Weibull distribution to the tail data and applying a hypothesis test to determine if the top score is an outlier with respect to the extrema of the non-match distribution. This prediction algorithm is functionally the same as the machine learning-based algorithm we explore in this work. The statistical model also provides a robust normalization [26] that produces match scores that can be combined using a typical fusion strategy (sum rule, product rule, etc.).

Finally, several works closely derived from the machine learning-based score analysis work of [16] can also be found in the literature. In [27], Wang et al. introduce a model of "perfect recognition" which treats the gallery as probe data, and computes a set of "perfect" similarity scores. A distance from each probe to this model can be computed and used as a feature for a machine learning-based prediction algorithm. This is a minor variation on the Δ features described in [16]. Further, for face recognition, Wang et al. use the same eye localization perturbation technique described in [11] to take corrective action based on the predictions made.

III. LEARNING FOR META-RECOGNITION

The goal of meta-recognition is to predict, in some automated fashion, if a match result is a success or failure. Statistical meta-recognition has several appealing characteristics that the machine learning-based algorithm does not have, including no need for training data, a straightforward implementation [10], and a provision for robust score normalization via the CDF of the chosen EVT distribution [26]. However, if we are able to train machine learning classifiers with the right combination of features, we can obtain more accurate results than with pure statistics for the prediction functionality. Here, the decisions made by the machine learning classifiers are the "predictions" in the theoretical meta-recognition framework. As part of our study in this article, we have developed novel ways to combine features for learning computed over raw recognition scores, as well as the decisions made by individual classifiers. In this section, we explore feature- and decision-level fusion to build enhanced classifiers for rank-1 recognition, and discuss the underlying operation of learning for meta-recognition. Complete source code implementing what is described here will be made available upon publication.

A. Features and Classifiers

1) Features from Scores: It is well-established that considering the raw scores from actual recognition algorithms as feature vectors for machine learning-based post-recognition score analysis, as we did with our statistical meta-recognition, generally does not work [11], [16], [27]. An explanation for this is provided in Sec. V. Thus, we define a set of features in accordance with [11]–[13], [16], [27]. We derive each feature from the distance measurements or similarity scores produced by the recognition algorithm. Before we calculate each feature, we sort the scores from best to worst. The top k scores $s_1, \ldots, s_k$ are used for the feature vector generation. We consider three different feature classes:

(a) $\Delta_{1,2}$ defined as $\langle s_1 - s_2 \rangle$. This is the separation between the top score and the second best score.

(b) $\Delta_{i,j...k}$ defined as $\langle(s_i - s_j), (s_i - s_{j+1}), \ldots, (s_i - s_k)\rangle$, where $j = i + 1$. Feature vectors may vary in length, as a function of the index i. For example, $\Delta_{1,2...k}$ is of length $k - 1$, $\Delta_{2,3...k}$ is of length $k - 2$, and $\Delta_{3,4...k}$ is of length $k - 3$.

(c) Discrete Cosine Transform (DCT) coefficients of the top-k scores. This is a variation on [11], where the Daubechies wavelet transform was shown to effectively represent the information contained in a score series.

Each feature considers the tail of a score distribution, in accordance with the theory of [10]. The rationale for the use of the Δ features is a shift invariance property, which prevents the learning from being impacted by the shifting of the distribution of the non-match scores as a function of the probe. If the underlying data exists in multiple scales (for example, multiple views in the gallery of the same person), then the DCT feature can account for this, as was shown for wavelet features in [11].

2) Building and Using Classifiers: First, we must collect the necessary training data to build a meta-recognition classifier. In order to build a classifier with balanced training and testing data, it is important to collect the same number of samples for both positive match instances (correct rank-1 recognition), and negative match instances (incorrect rank-1 recognition), from sequences of scores from the recognition algorithm for both. We use these scores as the source data for the features described in Sec. III-A1. The resulting feature vectors are tagged (positive or negative) and can be used as training input to any machine learning algorithm.

In this article, we have chosen support vector machines (SVM) just as an example; any supervised learning approach could also be used (indeed boosting [16] and neural networks [11] approaches have already been shown by our group to work for post-recognition score analysis). Our choice is motivated by the fact that SVMs are well known, do not depend on the dimensionality of the input space, use structural risk minimization, and allow fast testing (sub-linear time) of an input example. We evaluate the potential of other supervised learning approaches in Sec. IV-B.

From feature vectors, we can train an SVM classifier which seeks for learning the underlying nature of the score distributions. In practice, a radial basis function kernel yields the best results for this sort of feature data derived from scores. Linear and polynomial kernels were also tried, but did not produce results as accurate as the radial basis function kernel.

We do not need to train on a specific gallery (though this increases accuracy, as we show in Sec. V). Any probe/gallery combination is acceptable, as long as the training data does not overlap with the testing data for validation. This independence allows us to accommodate dynamic galleries. While the feature computation, as described in Sec. III-A1, does have a normalizing effect on the underlying data, it does not re-articulate the scores in a generalized manner. Existing schemes [11], [12], [16], [27], [28] have tended to train a classifier for each recognition algorithm or sensor being considered, based upon the scores from that algorithm only. During live recognition, we can compute a feature vector from the resulting scores, and simply apply the machine learning-based meta-recognition algorithm to make our prediction.

Given the above discussion, we can train an SVM classifier using Algorithm 1. For meta-recognition prediction, we use Algorithm 2. In each algorithm, Φ represents a feature.

Algorithm 1 Rank-1 Machine Learning (SVM) Training.

Require: A collection of similarity score sets $S_1^+, \ldots, S_n^+$. For each positive score set S_i^+ with z elements, the best score is a correct match.
Require: A collection of similarity score sets $S_1^-, \ldots, S_n^-$. For each negative score set S_i^- with z elements, the best score is an incorrect match.
1: **while** $i \leq n$ **do**
2:     **Sort** the scores, $s_1, \ldots, s_z \in S_i^+$;
3:     **Compute** Φ from Sec. III-A1 using $s_1, \ldots, s_n$; tag '+1'
4:     **Sort** the scores, $s_1, \ldots, s_z \in S_i^-$;
5:     **Compute** Φ from Sec. III-A1 using $s_1, \ldots, s_z$; tag '−1'
6:     $i \leftarrow i + 1$
7: **end while**
8: **Train** an SVM classifier using all $2n$ tagged feature vectors generating the classification model $\mathcal{M}_{\text{SVM}}$

Algorithm 2 Rank-1 Machine Learning (SVM) Meta-Recognition.

Require: A collection of similarity scores S with z elements.
Require: A classification model $\mathcal{M}_{\text{SVM}}$ trained from Algorithm 1.
1: **Sort** the scores, $s_1, \ldots, s_z \in S$;
2: **Compute** Φ from Sec. III-A1 using $s_1, \ldots, s_z$
3: **Classify** using $\mathcal{M}_{\text{SVM}}$
4: **if** *class-label* $c^* \geq 0$ **then**
5:     **Predict** Success
6: **else**
7:     **Predict** Failure
8: **end if**

B. Feature- and Decision-Level Fusion

We now turn to the description of each fusion technique we use to enhance the prediction accuracy of machine learning-based meta-recognition. Considering lower levels within the learning framework, we can fuse the recognition algorithm scores before computing the features described in Sec. III-A1 in a blending approach similar to [14]. Similarly, different computed features can be concatenated into a single feature vector to provide more information to the learning.

We can also consider higher levels of the system. Decision-level fusion [4] is defined as data processing by independent classifiers, followed by the fusion of decisions (based upon the calculated results) of each classifier. This idea can be thought of as n different inputs to n different classifiers, producing n decisions that are combined together to produce a final decision that the system will act upon. Motivating this, decision-level fusion via thresholding has been shown to be quite effective in a biometric classification context [29].

The choice of decision-level fusion for meta-recognition stems from our need to combine data over independent recognition algorithms or sensors, as well as decisions from multiple independent classifiers. In the following descriptions for each specific fusion technique, $\mathcal{T}$ is a threshold, Φ is one of the features in Sec. III-A1, and s_i is a score from a recognition algorithm.

(a) Multi-algorithm Fusion.
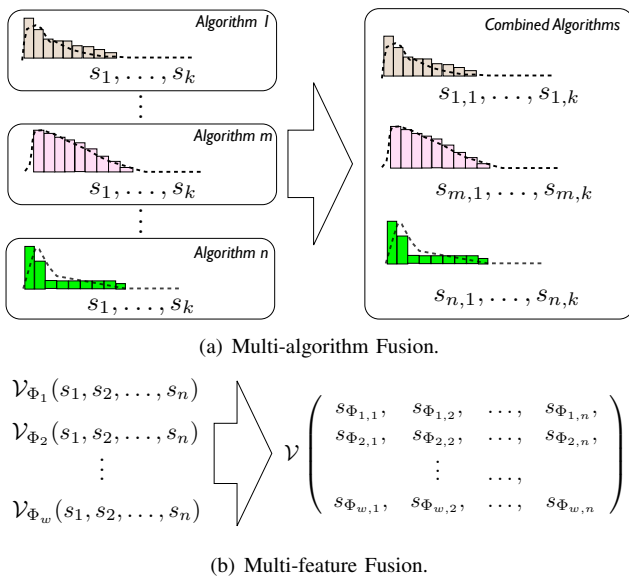
(b) Multi-feature Fusion.

Fig. 3. Illustrations representing combination oriented fusion approaches.

- Combine data from one or more recognition algorithms (Fig. 3(a)): Some scores from one or more recognition algorithms are combined with the scores from another recognition algorithm to improve the accuracy of classification. Fusion here takes place before score feature generation for training or classification, with one feature Φ applied to the scores from each recognition algorithm in the combined data.

- Consider a combination of different score features (Fig. 3(b)): This technique combines the feature vectors that have been calculated for individual features before training or classification into one feature vector. This blending is a concatenation meant to lift the performance in the machine learning by enhancing classification with longer, and ideally more distinct, feature vectors.

- Threshold over all decisions D from machine learning meta-recognition classifiers across features: $\mathcal{T}(D(\Phi_1), D(\Phi_2), \ldots, D(\Phi_w))$. With this technique, we set a single threshold over classification decisions across features for a single algorithm, or for classification decisions across algorithms. If k decisions exceed $\mathcal{T}$, where $k \geq 1$, then predict "success".

- Individual thresholds across all decisions across score features: $(\mathcal{T}_1(D(\Phi_1)), \mathcal{T}_2(D(\Phi_2)), \ldots, \mathcal{T}_w(D(\Phi_w)))$. With this technique, we set individual thresholds for each classification decision across different features for a single algorithm, or for classification decisions across algorithms. If k decisions exceed $\mathcal{T}_i$, where $k \geq 1$, then predict "success".

Missing data impacts these fusion techniques in two ways. First, in the case of the feature-level fusion technique that combines data from several recognition algorithms, if data for an entire algorithm or sensor is missing, we end up with an uneven feature vector, which strongly impacts the accuracy of our learning. Second, in the case of all of these techniques, if we do not have at least 20 available scores (the longest feature we consider in this work is $\Delta_{1,2,\ldots 10}$) from a matching instance for feature generation, we violate the theory of meta-recognition. If only a modest amount of data is missing (holes at various scattered ranks, for instance), we are still able to sample from the tail of the overall score distribution for feature generation, with little to no impact on the meta-recognition accuracy. Our strategy to handle severe cases of missing data in this work is to simply "fail over" to a technique where sufficient data exists.

IV. MACHINE LEARNING META-RECOGNITION RESULTS

In this section, we present a validation of our machine learning-based algorithm for meta-recognition, comparing baseline features to our enhanced fusion-oriented classifiers, along with the experimental framework used for all assessments. Our goal is to evaluate, in an empirical manner, all features and fusion techniques that are defined in Sec. III.

A. Meta-Recognition Error Trade-off Curves

All source data used for the experiments in this article are scores from identification instances. Our goal is to evaluate our machine learning algorithm for meta-recognition, as well as other comparison methods that produce a prediction of recognition *success* or *failure*. In order to assess the accuracy of meta-recognition predictions, 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.

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

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

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

$$MRFAR = \frac{|C_1|}{|C_1| + |C_4|}, \quad MRMDR = \frac{|C_2|}{|C_2| + |C_3|}. \quad (1)$$

This representation is a convenient indication of meta-recognition accuracy, and we use it to express all the results we present in this article. The MRFAR and MRMDR can be adjusted via thresholding applied to the predictions (for the machine learning experiments in this article, these predictions are SVM decision scores), to build the curve. Just as one uses a traditional DET or ROC curve to set recognition system
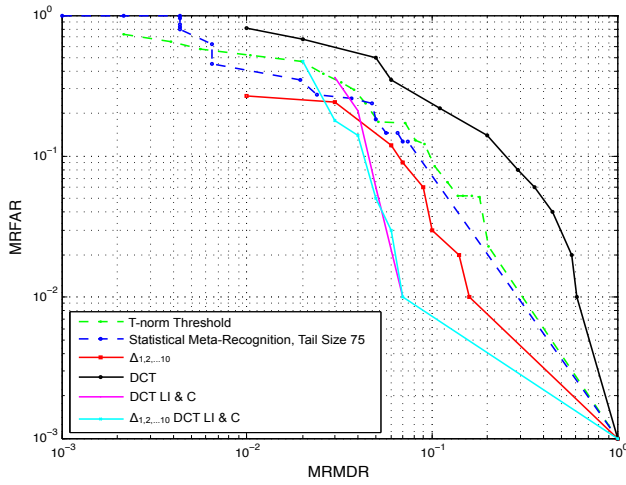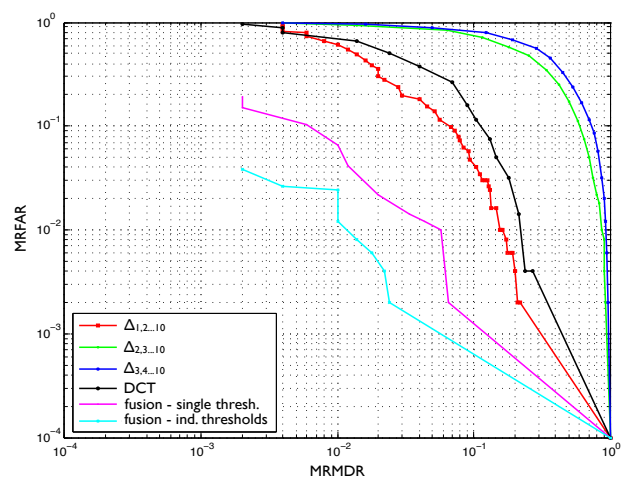
(a) BSSR1 Face Algorithm C
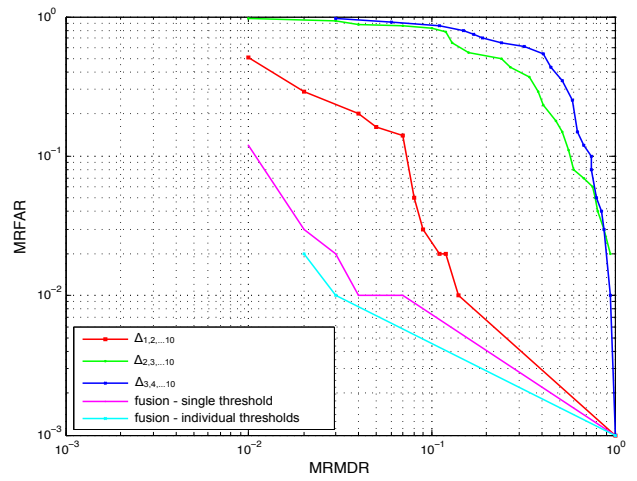
(a) BSSR1 Face Algorithm G
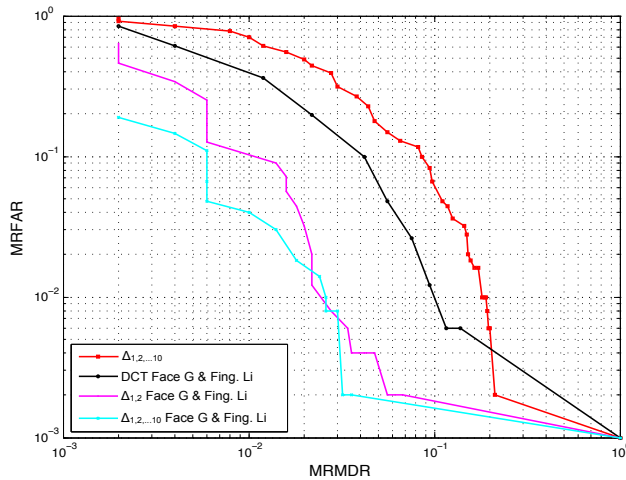
(b) BSSR1 "Chimera" Face Algorithm G

(b) BSSR1 "Chimera" Face Algorithm C

Fig. 4. MRET curves for multiple biometric recognition algorithms from the BSSR1 data set generated using a machine learning-based meta-recognition system. Each plot also depicts mixed modality algorithm blending fusion for the best performing features calculated over the scores of BSSR1 Alg. LI. Comparison curves for basic thresholding over T-norm scores and statistical meta-recognition are shown in (a). Note that in (b), the $\Delta_{1,2,\ldots 10}$ feature is the best performing baseline, while the rest of the curves reflect the fusion of multiple sources of scores.

Fig. 5. MRET curves for multiple biometric recognition systems from the BSSR1 data set generated using a machine learning-based meta-recognition system. Each plot depicts enhanced meta-recognition with single threshold and multiple threshold fusion. The application of these decision-level fusion techniques significantly enhances prediction accuracy.

parameters, the meta-recognition parameters can be tuned using the MRET.

B. Experimental Meta-Recognition

Our experimental evaluation of the machine learning-based meta-recognition algorithm has two primary goals in mind: first, to show the accuracy advantage of the fusion techniques over the baseline features for meta-recognition; and second, to show the accuracy advantage of the machine learning-based meta-recognition algorithm over the statistical meta-recognition algorithm and a cohort model of post-recognition score analysis. These experiments represent the operation of the Y component in Def. 1.

Table I shows the data used for experimentation, including training and testing breakdowns, as well as the specific

recognition algorithms considered. We note that the gallery considered in all experiments is always consistent between training and testing data. Only the probes vary to produce a different score series for every feature vector. This methodology is consistent with the standard usage of these data sets.

For the first round of experiments, the NIST multi-biometric BSSR1 data set was used. The subset of this data set that provides true multi-biometric results (fing_x_face) is relatively small for a learning test, providing match scores for 517 unique probes across two face recognition algorithms (labeled C & G), and scores for two fingers (labeled LI & RI) for one fingerprint recognition algorithm. In order to gather enough negative data for training and testing, negative examples for each score set were generated by removing the top score from matching examples. In order to address the limited nature of the multi-biometric BSSR1 set, we created a "chimera" data set from the larger face and finger subsets provided by BSSR1 (3,000 score sets each for two face algorithms, and 6,000

| Data Set	Training Samples	Test Samples	Recog. Algs.
BSSR1 [30]	600	200	2 Face & 1 Finger
BSSR1 "chimera"	6000	1000	2 Face & 1 Finger
GBU [31]	200	200	LRPCA
ALOI [32]	200	180	SIFT
"Corel Relevants" [33]	300	200	4 CBIR

TABLE I
DATA BREAKDOWN FOR MACHINE LEARNING META-RECOGNITION.
TESTING AND TRAINING DATA IS PER ALGORITHM (SOME SETS CONTAIN
MORE THAN 1 ALGORITHM). NO OVERLAP EXISTS BETWEEN SETS.

score sets each for two sampled fingers for a single fingerprint algorithm), which are not inherently consistent across scores for a single user. This chimera set is artificially consistent across scores for a single user, and provides us with much more data to consider for fusion.

Results for a selection of data across both the true multi-biometric and chimera sets, all algorithms, are presented as MRET curves in Figs. 4 and 5. To interpret these plots (and the following MRET curves found throughout the rest of this article), it must be understood that points approaching the lower left corner minimize both the MRFAR and MRMDR errors. Only features that produced results better than chance are plotted. Algorithm blending fusion across modalities (Fig. 4), as well as single threshold fusion and individual thresholds fusion (Fig. 5), improve the performance of meta-recognition, compared with the baseline features, as well as statistical meta-recognition [10] (an example comparison is shown in Fig. 4 (a)). Both the machine learning-based algorithm and the statistical algorithm use exactly the same score data as input for each comparison test, making this a fair comparison. Feature blending fusion (not plotted) produced results as good as the best performing feature, but never significantly better. Different combinations of blending were attempted including mixing all features together, as well as different subsets of the features. While not improving meta-recognition accuracy, this fusion technique predicts performance and implicitly indicates the best performing feature, without prior knowledge of the accuracy produced by any particular feature.

We also confirmed that the machine learning-based meta-recognition algorithm is significantly better than a standard threshold test over T-norm scores [24] [25]. T-norm scores were generated, following [24], 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 $(|non\text{-}match\ distribution| - 100)$ 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. 4(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.
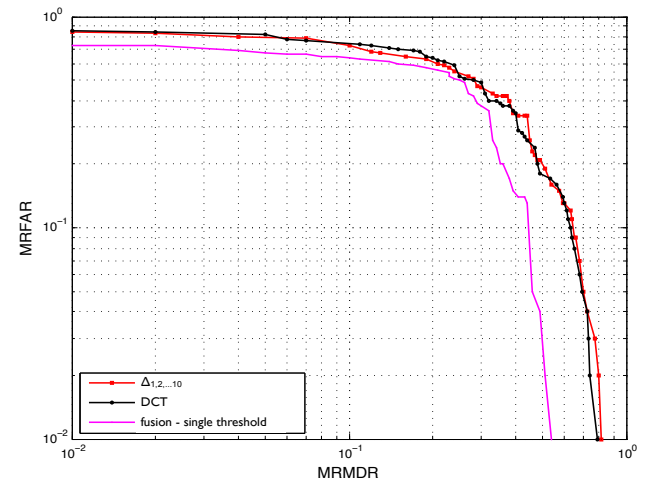
Fig. 4(a) shows that in every case except the baseline DCT feature, the meta-recognition technique significantly outperforms the T-norm based thresholding (labeled T-norm Threshold). The equal error rate (the point at which MRFAR
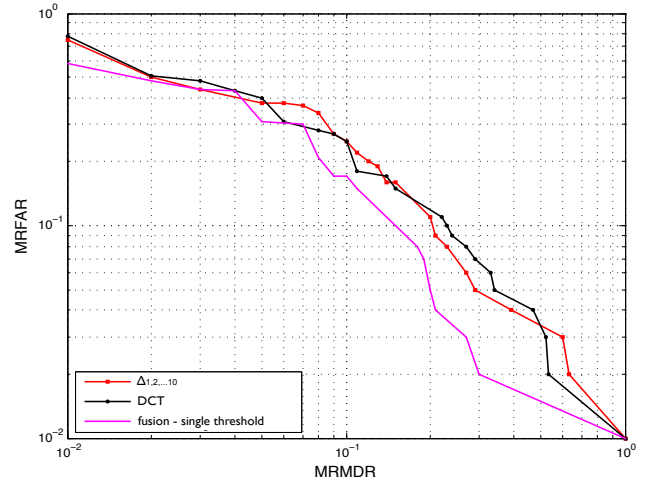
(a) Good
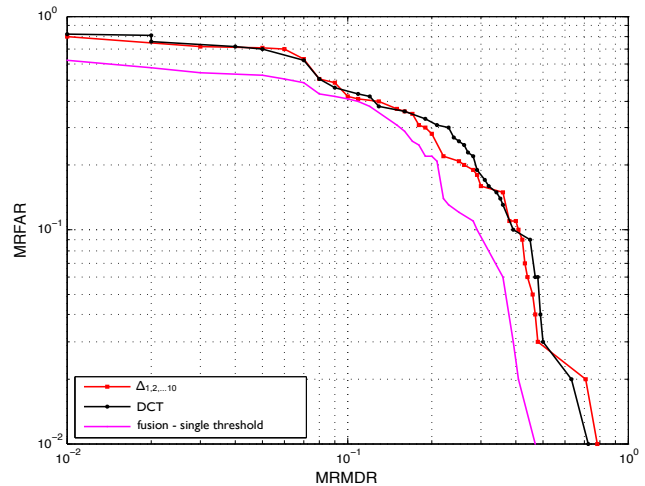
(b) Bad

(c) Ugly

Fig. 6. MRET curves for the Good, the Bad, & the Ugly face recognition challenge problem. Each plot shows results for the same two baseline features, as well as single threshold fusion over those features. There is a distinct difference in prediction accuracy between each partition of the GBU set, with Good (a) producing the most accurate results, and Ugly (c) producing the least accurate results.

and MRMDR errors are equal) for the $\Delta_{1,2,...10}$ LI & C curve is roughly 5%, meaning that just 1 out of 20 meta-recognition instances will incorrectly predict success or failure for this algorithm and this classifier. The T-norm Threshold curve has an equal error rate of 10%, and is much worse at other points along the curve in comparison to the other MRET curves for meta-recognition. This is consistent with the findings of [10].

While BSSR1 provided useful data for our analysis, it is an older set and does not provide a large amount of score variation. Our second round of experiments considered data from the Good, the Bad, and the Ugly face recognition challenge problem [31], the most recent large scale face set available from the University of Notre Dame. This set provides three partitions of the well-known MBGC [34] set that are organized by difficulty. Using the CSU LRPCA baseline algorithm[2], we calculated match scores for each partition. With a need for both training and testing data from the same pool of scores, we were bounded by the partition with the lowest rank-1 recognition rate: Ugly. Thus, a total of 200 score vectors were sampled for training, and 200 were sampled for testing for each partition (with negative score vectors generated, as described above, when needed). For a fair comparison between sets, we did not alter the training and testing counts between Good, Bad and Ugly. In total, however, this data consists of 1,302,000 scores (651,000 scores each for training and testing).

For each partition, we computed the best performing features noted in the curves from the first round of experiments, $\Delta_{1,2,...10}$ and DCT, and then fused them using the individual threshold approach. The results of our GBU experiments are shown in the plots of Fig. 6. As expected, the individual threshold fusion enhanced the prediction accuracy in each case. More interesting though is the comparison between sets. We observe a distinct difference in prediction accuracy between each partition of the GBU set, with Good (Fig. 6 (a)) producing the most accurate results, and Ugly (Fig. 6 (c)) producing the least accurate results. This observation is consistent with our theory; as the score distributions lose statistical variation with increased difficulty, the positive and negative features vectors start to become ambiguous. In essence, the extrema almost become confusers (outliers in their own right) with respect to the correct match, and the SVM learning has trouble discerning the distinction.

For the third round of experiments, we tested a series of popular object recognition algorithms that can be used for digital image forensics applications using the machine learning-based meta-recognition algorithm. In this case, we have considered standard SIFT [35] with all features except DCT, which did not yield results better than random chance for our data. For the content-based image retrieval experiments, we considered simple local color histograms. All features except for $\Delta_{3,4,...10}$ were utilized for CBIR. The results of Fig. 7 show a significant increase in accuracy for the fusion techniques.

When considering the accuracy of individual features, we can see from our results that the $\Delta_{1,2,...10}$ feature is consistently the best, while the DCT and $\Delta_{1,2}$ features sometimes

[2]http://www.cs.colostate.edu/facerec/algorithms/lrpca2010.php
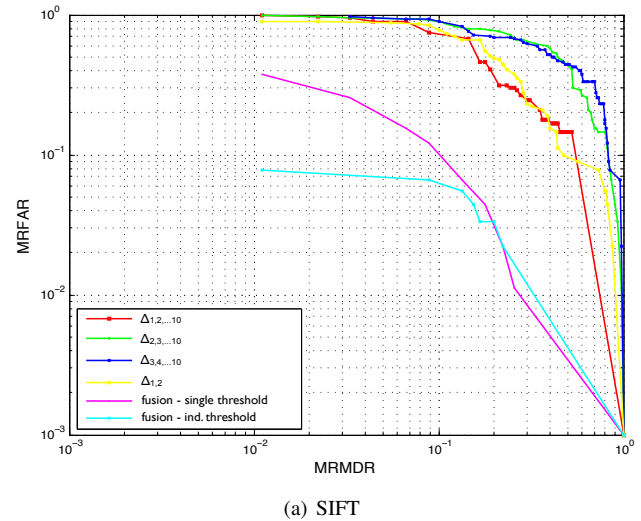
(a) SIFT
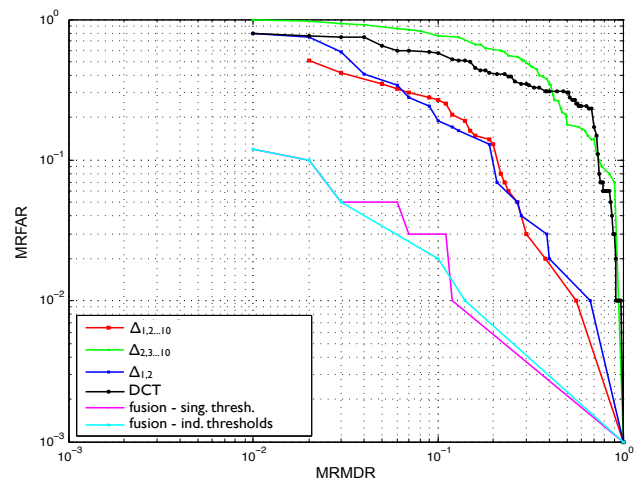
(b) CBIR Algorithm LCH

Fig. 7. MRET curves for multiple object recognition systems generated using a machine learning meta-recognition system. Each plot depicts enhanced meta-recognition with single threshold and individual thresholds fusion.

produce good accuracy, and sometimes do not. As individual features, $\Delta_{2,3,...10}$ and $\Delta_{3,4,...10}$ produce the least accurate results. When considering the feature-level algorithm blending fusion and decision-level single threshold and individual thresholds fusion approaches, the prediction results for all recognition algorithms are significantly enhanced, well beyond the baseline features. Thus, these fusion approaches produce the best meta-recognition results observed in all of our experimentation. Since the cost to compute multiple features is negligible, feature- and decision-level fusion can easily be run for each meta-recognition attempt in an operational setting.

Finally, coming back to the discussion in Sec. III-A2 on the potential of other supervised learning approaches, we took a look at other algorithms besides SVM. Machine learning meta-recognition classifiers were trained for the following algorithms [36]: AdaBoost, Bootstrap Aggregating (Bagging), Linear Discriminant Analysis (LDA), and Neural Networks. The features used for training and testing were exactly the same as those used for the BSSR1 "Chimera" Face Algorithm
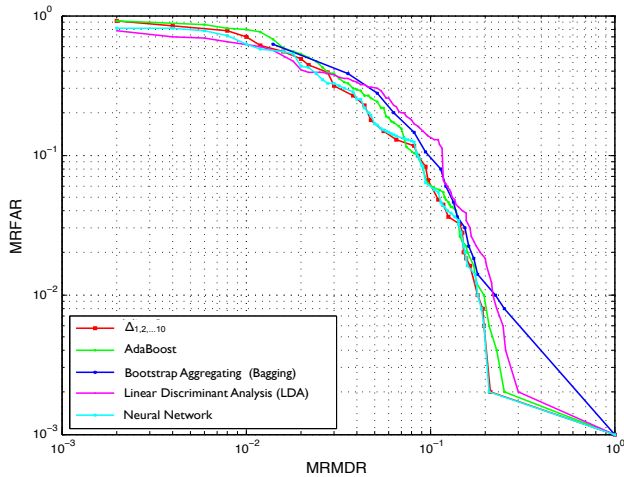
Fig. 8. MRET curves for the BSSR1 "Chimera" Face Algorithm G data generated using different machine learning-based meta-recognition classifiers. As described in Sec. III-A2, any supervised learning approach can be used for meta-recognition. These curves, which show that all of the evaluated approaches are close in accuracy, provide evidence for that claim. This experiment allows us to conclude that it is the features capturing the shape of the extreme scores, which provides for effective machine-learning Meta-recognition.
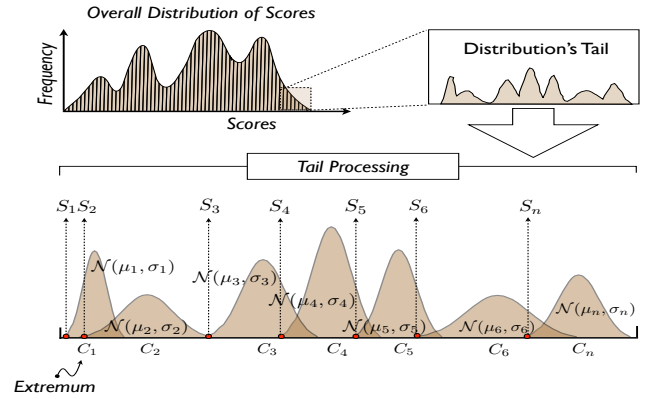

Fig. 9. The data generation process for an experiment with synthetic data comparing statistical and machine learning-based meta-recognition. Our goal is to produce a realistic sampling of "scores" that conforms to the portfolio model described in [10]. A series of Gaussian distributions are generated to represent different gallery classes, each with a randomly generated mean and standard deviation to force intra-class dependency (bottom of figure). From each of these distributions, the extremum ("best score") is selected, and placed into a sorted sequence. This sequence represents the tail of the overall distribution of scores, which we can use for machine learning feature generation or Weibull fitting.

G $\Delta_{1,2,...10}$ feature in Fig. 4. The procedure for using these other approaches follows what is described in Algs. 1 and 2, with the exception that the classification model $\mathcal{M}$ is not constrained to just SVM. We trained the selected classifiers using their standard out-of-the-box configurations and parameters. No parameter tuning was performed. From Fig. 8, it can be seen that the prediction accuracies for all of the learning approaches (including SVM) are rather close. This outcome is not surprising, given that no single supervised learning approach has been shown to be superior with respect to classification accuracy in the machine learning literature. In essence, all of these algorithms are suitable for machine learning-based meta-recognition, handing the decision of which one to use to the interested user.

V. PURE STATISTICS VS. MACHINE LEARNING

Several experiments in this work correspond to existing experiments for statistical meta-recognition in [10] (Figs. 6(b), 7(a) and 7(b) in [10] vs. Figs. 4(a), 5(a), 7(a), and 7(b) in this work). In all of these cases, the machine learning-based algorithm with feature- or decision-level fusion produces more accurate results. Moreover, we also find instances such as the one in Fig. 4(a), where baseline machine learning-based meta-recognition features produce better meta-recognition results when compared to the pure statistical algorithm. To account for this performance improvement, we consider the most significant differences between the machine learning and statistical algorithms – namely, the use of features computed from scores, and the use of training data by the machine learning algorithm.

We would like to verify that the features of Sec. III-A1 have a normalizing effect upon the data they are applied to. As discussed in our previous work [10], the Generalized Extreme Value distribution is a 3-parameter family: one parameter

shifting its location, one its scale and one that changes its shape. The EVT theory provides the reason why the learning algorithm is successful. The learning can develop an implicit overall Weibull shape parameter, ignoring any shift since the learning features are shift-invariant, and test the outlier hypothesis effectively. The failure of the learning algorithm on the raw data is likely caused by the shifting of the distribution of the non-match scores as a function of the probe. The operation of our learning algorithm, where we consider an n-element feature space composed of k-dimensional feature data from matching and non-matching scores, is just a corollary to the EVT, adapted to the recognition problem.

However, normalization by itself should not produce any significant improvement in accuracy. We expect the results, in the best case, to be as good as statistical meta-recognition. Thus, we turn our attention to the effects of training data. It is apparent that with machine learning, the meta-recognition classifiers have knowledge of numerous training examples, whereas there is no *a priori* knowledge for statistical classifiers. Statistical meta-recognition only considers the scores for the recognition instance at hand. As noted in Section IV, the experiments presented thus far have always considered the same gallery for training and testing, with only the probe varying to produce different score sequences. If the machine learning is able to learn gallery information from the training samples and apply that knowledge to the classification of test samples derived from the same gallery, then the gallery's influence on the learning accuracy is an important contributing factor to its advantage.

To explore this claim, we designed a large-scale experiment with synthetic data reflecting the portfolio model found in our previous work [10]. With synthetic data, we are able to control for data dependency and sampling, thus eliminating additional variables that affect the applicability of the statistical theory, but are not immediately relevant to this analysis. With a

substantially larger statistically relevant sampling of scores, we also eliminated any small sampling effects that might have been present in the BSSR1 experiments of Section IV, where we were constrained by a well-known, but limited data set.

In order to evaluate our hypothesis that a consistent gallery between training and testing gives machine learning based meta-recognition an advantage, three different score sets were generated using the process described in Fig. 9: training sets R_1, R_2 and testing set T_1. The sets R_1 and T_1 have 150 sets of distance scores, with no overlap in their probes, and share a common gallery. R_2 has 150 sets of distance scores with no overlap in probe or gallery. The R_1 and T_1 score sets have a total of 4,455,000 scores per set. The R_2 score set has a total of 3,712,500 scores for the set.

The statistical algorithm for this experiment utilized a tail size of five scores for fitting (out of numerous experiments, a size of five yielded the most accurate results). For the machine learning classifiers, a $\Delta_{1,2...5}$ feature was used for training and testing (the same score data used by the statistical algorithm). The first machine learning classifier was trained with R_1 while the second machine learning classifier for the comparison was trained with R_2. The results in Fig. 10 support our hypothesis. When the training gallery does not overlap with the testing gallery (R_2 for training and T_1 for testing, represented by the black curve), the prediction accuracy is quite consistent with the accuracy of statistical meta-recognition (represented by the green curve). When the training gallery is the same as the testing gallery (R_1 and T_1, represented by the red curve), there is a noticeable increase in prediction accuracy. The operational impact of this finding is clear. If a recognition system incorporating meta-recognition wishes to achieve the highest levels of prediction accuracy that are currently known to be possible, then the system should be designed with a machine learning-based classifier trained on the same gallery that will be used during operation, and ideally use feature- or decision-level fusion.

VI. Conclusion

With two different options to implement meta-recognition, the question of which one to choose for an operational scenario is an important one addressed in our study. Statistical meta-recognition achieves accuracies far beyond standard thresholding and cohort thresholding, without the need for training data. However, our machine learning-based algorithm for meta-recognition has provided us with more accurate results than the statistical algorithm over many experiments.

In this article we have sought a deeper understanding of the underlying feature mechanisms for the machine learning-based algorithm that lead to higher accuracies through feature- and decision-level fusion. Further, we were also able to show that a consistent gallery between training and testing data increases the accuracy of machine learning-based meta-recognition – even without feature- or decision-level fusion. An observation from this work that should be of general interest to researches is the need to investigate possible biases in learning based approaches. In this case, the bias turned out to be a valid way to organize our data. In other cases, biases may lead to higher levels of classification accuracy than can achieved in practice.

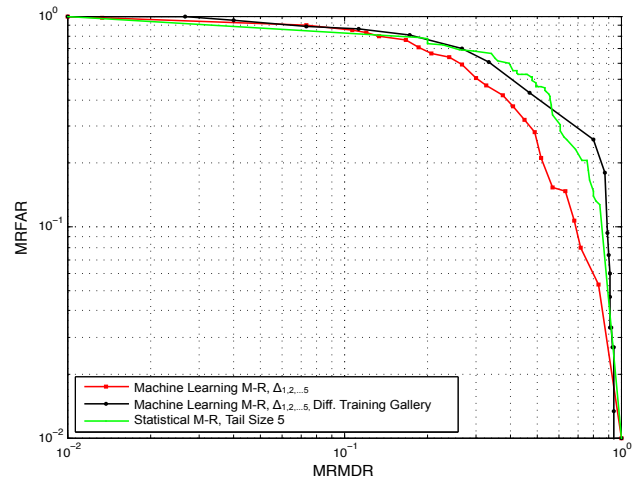Fig. 10. MRET curves showing Weibull-based statistical meta-recognition accuracy [10] *vs.* machine learning-based meta-recognition accuracy for a controlled synthetic data set reflecting the portfolio model. The machine learning-based algorithm shows a clear advantage over the pure statistical algorithm when a consistent gallery is considered for training and testing. When the machine learning classifier is trained and tested using different galleries, performance is approximate to that of the statistical algorithm.

Consideration for operational scenarios continues to drive this work forward in several ways. Most importantly, we are investigating new algorithms to extend the prediction capabilities of machine learning-based meta-recognition beyond rank-1 to rank-n. This presents several challenges for SVM learning. If we approach this as a multi-class problem, we may lose some of the discriminative power we have with binary classifiers. Similarly, the issue noted in this article of extrema almost becoming confusers to the learning with respect to the correct match is exacerbated at higher ranks. Thus, supervised learning algorithms other than SVM, such as the ones we looked at in this article, are being considered. Finally, we are exploring the specific potential of meta-recognition for cases of plastic surgery and disguise, so we can move towards automatically identifying a new crop of savvy outlaws such as Ramírez Abadía.

Acknowledgment

References

[1] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary Face Recognition Grand Challenge Results," in *Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 15–24.

[2] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni, "Fingerprint Verification Competition 2006," *Biometric Technology Today*, vol. 15, no. 7-8, pp. 7–9, 2007.

[3] N. Pinto and D. Cox, "Beyond Simple Features: a Large-Scale Feature Search Approach to Unconstrained Face Recognition," in *IEEE AFGR*, March 2011.

[4] A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*. Springer, 2006.

[5] The Guardian, "My Boy Lollipop: Raid Ends Sweet Life of the Colombian Drug Lord," August 11 2007, http://www.guardian.co.uk/world/2007/aug/11/colombia.brazil.

[6] BBC News, "'Fake Fingerprint' Chinese Woman Fools Japan Controls," December 2009, http://news.bbc.co.uk/2/hi/asia-pacific/8400222.stm.

[7] N. Ramanathan, R. Chellapa, and A. K. R. Chawdhury, "Facial Similarity Across Age, Disguise, Illumination, and Pose," in *IEEE Intl. Conf. on Image Processing*, 2011, pp. 1999–2002.

[8] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyezdan, "Plastic Surgery: A New Dimension to Face Recognition," *IEEE Trans. on Inf. Forensics and Security*, vol. 5, pp. 441–448, 2010.

[9] Agence France-Presse, "Hello Kitty Used as Drug Lord's Messenger: Report," March 10 2008, http://afp.google.com/article/ALeqM5ieuIvbrvmfofmOt8o0YfXzbysVuQ.

[10] W. Scheirer, A. Rocha, R. Michaels, and T. E. Boult, "Meta-Recognition: The Theory and Practice of Recognition Score Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1689–1695, 2011.

[11] T. Riopka and T. Boult, "Classification Enhancement via Biometric Pattern Perturbation," in *IAPR Intl. Conf. on Audio- and Video-Based Biometric Person Authentication*, vol. 3546, 2005, pp. 850–859.

[12] W. Scheirer, A. Bendale, and T. Boult, "Predicting Biometric Facial Recognition Failure With Similarity Surfaces and Support Vector Machines," in *IEEE Intl. Workshop on Biometrics*, 2008.

[13] W. Scheirer and T. Boult, "A Fusion-Based Approach to Enhancing Multi-Modal Biometric Recognition System Failure Prediction and Overall Performance," in *IEEE Intl. Conf. on Biometrics Theory, Applications and Systems*, 2008.

[14] B. Xie, T. Boult, V. Ramesh, and Y. Zhu, "Multi-Camera Face Recognition by Reliability-Based Selection," in *IEEE Intl. Conf. on Computational Intelligence for Homeland Security and Personal Safety*, 2006.

[15] B. Xie, V. Ramesh, Y. Zhu, and T. Boult, "On Channel Reliability Measure Training for Multi-Camera Face Recognition," in *IEEE Intl. Workshop on Applications of Computer Vision*, 2007.

[16] W. Li, X. Gao, and T. Boult, "Predicting Biometric System Failure," in *IEEE Intl. Conf. on Computational Intelligence for Homeland Security and Personal Safety*, 2005.

[17] R. Duda, P. Hart, and D. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[18] C. Bishop, *Neural Network for Pattern Recognition*. Clarendon Press, 1995.

[19] P. Grother and E. Tabassi, "Performance of Biometric Quality Evaluations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

[20] 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.

[21] P. Phillips and J. R. Beveridge, "An Introduction to Biometric-completeness: The Equivalence of Matching and Quality," in *IEEE Intl. Conf. on Biometrics Theory, Applications and Systems*, 2009.

[22] G. Aggarwal, N. Ratha, R. Bolle, and R. Chellappa, "Multi-biometric Cohort Analysis for Biometric Fusion," in *Intl. Conf. on Acoustics, Speech and Signal Processing*, 2008.

[23] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[24] N. Poh, A. Merati, and J. Kittler, "Adaptive Client-Impostor Centric Score Normalization: A Case Study in Fingerprint Verication," in *IEEE Intl. Conf. on Biometrics Theory, Applications and Systems*, 2009.

[25] ——, "Making Better Biometric Decisions with Quality and Cohort Information: A Case Study in Fingerprint Verication," in *Eurasip European Signal Processing Conf*, 2009.

[26] W. Scheirer, A. Rocha, R. Michaels, and T. E. Boult, "Extreme Value Theory for Recognition Score Normalization," in *European Conf. on Computer Vision*, 2010.

[27] P. Wang, Q. Ji, and J. Wayman, "Modeling and Predicting Face Recognition System Performance Based on Analysis of Similarity Scores," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 665–670, 2007.

[28] P. Wang and Q. Ji, "Performance Modeling and Prediction of Face Recognition Systems," in *Intl. Conf. on Computer Vision and Pattern Recognition*, 2006.

[29] K. Veeramachaneni, L. Osadciw, A. Ross, and N. Srinivas, "Decision-level Fusion Strategies for Correlated Biometric Classifiers," in *IEEE Intl. Workshop on Biometrics*, 2008.

[30] N. I. of Standards and Technology, "NIST Biometric Scores Set," 2004, http://www.itl.nist.gov/iad/894.03/biometricscores/.

[31] P. Phillips, J. R. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. Lui, H. Sahibzada, and S. Weimer, "An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem," in *IEEE AFGR*, March 2011.

[32] J. Geusebroek, G. Burghouts, and A. Smeulders, "The Amsterdam Library of Object Images," *Intl. Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[33] R. Stehling, M. Nascimento, and A. Falcão, "A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification," in *ACM Intl. Conf. on Inf. and Knowledge Management*, 2002, pp. 102–109.

[34] P. Phillips, J. Beveridge, W. Scruggs, A. O'Toole, D. Bolme, K. Bowyer, B. Draper, G. Givens, Y. M. Lui, H. Sahibzada, J. S. III, , and S. Weimer, "Overview of the Multiple Biometrics Grand Challenge," in *Proc. of ICB*, 2009.

[35] D. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[36] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

Walter J. Scheirer received his M.S. degree in computer science from Lehigh University (2006), and his Ph.D. in Engineering, from the University of Colorado at Colorado Springs (2009). He is currently the director of research and development at Securics, Inc., and an assistant professor adjoint at the University of Colorado at Colorado Springs. His primary research interests include biometrics, computer vision, digital image forensics, machine learning, and digital humanities.

Anderson de Rezende Rocha received his CS B.Sc degree from Federal University of Lavras (UFLA), Brazil in 2003. He received his Computer Science M.S. and Ph.D. from University of Campinas (Unicamp), Brazil, in 2006 and 2009, respectively. Currently, he is an assistant professor in the Institute of Computing, Unicamp. As of 2011, Prof. Rocha is a Microsoft Research Faculty Fellow and an elected member of the Brazilian Academy of Sciences. He is also an elected member of the IEEE IFS-TC. His interests include digital image and video forensics, machine intelligence, and general computer vision.

Jonathan Parris received his B.S. in Computer Science from the University of Colorado at Colorado Springs. He is currently working on his M.S in Computer Science at UCCS. He works as a research assistant in the Vision and Security Technology Lab at UCCS. His research interests are in computer vision and machine learning.

Terrance E. Boult , El Pomar Professor of Innovation and Security at the University of Colorado at Colorado Springs had published over 180 Papers and holds 9 patents (8 pending). Prior to joining UCCS, Dr. Boult held professorships at Lehigh and Columbia universities. He is also CEO/CTO of Securics Inc., a company in the biometrics and security space. Dr. Boult has served as an Assoc. Editor for TPAMI, has been the PAMI-TC chair and is a member of the IEEE Golden Core Society.