Robust Fusion: Extreme Value Theory for Recognition Score Normalization

Walter Scheirer[1], Anderson Rocha[2], Ross Micheals[3], and Terrance Boult[1]*

[1] University of Colorado at Colorado Springs & Securics, Inc.
[2] Institute of Computing, University of Campinas
[3] National Institute of Standards and Technology

Abstract. Recognition problems in computer vision often benefit from a fusion of different algorithms and/or sensors, with score level fusion being among the most widely used fusion approaches. Choosing an appropriate score normalization technique before fusion is a fundamentally difficult problem because of the disparate nature of the underlying distributions of scores for different sources of data. Further complications are introduced when one or more fusion inputs outright fail or have adversarial inputs, which we find in the fields of biometrics and forgery detection. Ideally a score normalization should be robust to model assumptions, modeling errors, and parameter estimation errors, as well as robust to algorithm failure. In this paper, we introduce the w-score, a new technique for robust recognition score normalization. We do not assume a match or non-match distribution, but instead suggest that the top scores of a recognition system's non-match scores follow the statistical Extreme Value Theory, and show how to use that to provide consistent robust normalization with a strong statistical basis.

1 Introduction

For many different recognition problems in computer vision, the ability to combine the results of multiple algorithms and/or sensors brings significant improvement in overall recognition performance. While there are many approaches and "levels" of fusion, a widely used approach is score level fusion, where scores from different recognition algorithms are combined. Since score distributions vary as a function of the recognition algorithms, and sometimes the underlying sensors, one must normalize the score data before combining it in score level fusion.

The goal of fusion is to improve recognition accuracy, and hence it is important that the underlying process be robust. Choosing a robust score normalization technique is often a challenge for several reasons. In the literature, the term *robust* has been defined as insensitivity to the presence of outliers (noise) [1] for the estimation of any necessary parameters. While this definition captures one property of good fusion, there are more issues than just the parameter estimation. We define the term *robust fusion* to be a fusion process (including

* This work was supported by ONR STTR N00014-07-M-0421, ONR SBIR N00014-09-M-0448, ONR MURI N00014-08-1-0638 and Fapesp Grant 2008/08681-9.

normalization) that is insensitive to errors in its distributional assumptions on the data, has simple parameter estimation, and a high input failure tolerance. For this work, simple parameter estimation means there is no dependence on a large sample set for modeling the match and non-match distributions for each algorithm, and a very small number of parameters must be estimated experimentally. Failure tolerance means that if one or more recognition algorithms involved in the fusion process is not producing correct matching results, it does not strongly impact the final result of fusion. Ideally, we would like a score normalization that is both robust to failure, and is unencumbered by complicated parameter estimation as score distributions vary. Further, if an algorithm is repeatedly failing, robust fusion should be able to detect this.

Robustness in score level fusion is strongly impacted by normalization in two major ways:

1. The varying nature of the underlying distribution of scores across different recognition algorithms often leads to inconsistency in normalization results. For example, if a normalization technique assumes the algorithms considered for fusion produce scores that follow a Gaussian distribution, and at least one of those distributions is not Gaussian, the results will not be optimal. The distribution of recognition scores is the result of a complex function of both the algorithm and the actual data being processed, and it is dangerous to assume too much about the score distribution.
2. Complications are introduced when one or more sensors or recognition algorithms being considered for fusion fail or are deceived. For recognition problems, failure occurs when an input sample of a class unknown to the system is recognized as being part of a known class, or when an input sample that should be recognized by the system is rejected as being unknown. The scores produced in these failure scenarios become problematic for normalization techniques, especially when they resemble an "expected" (and often estimated) match distribution.

In this paper, we introduce a new score normalization approach for robust fusion based on a probability of confidence that a particular score is not drawn from the non-match distribution. For an overview, we turn to Figure 1. Based on the match scores produced by multiple recognition algorithms applied to a particular object, a post-recognition score analysis [2] [3] is performed to predict the probability of the scores not being from the non-match distribution. For this work, we introduce a statistical Extreme Value Theory normalization that draws these probabilities from the cumulative distribution function of a Weibull distribution (hence "w-score"). The resulting probabilities from the different algorithms are the normalized w-scores, which can then be fused together to produce an overall probability of not being a non-match. In Figure 1, the process is shown for the case of two algorithms, though it applies to any number of inputs.

Traditional normalization techniques change the location and scale parameters of a score distribution in an ad-hoc manner or based on unproven distributional assumption. In contrast, our w-score normalization changes raw scores to

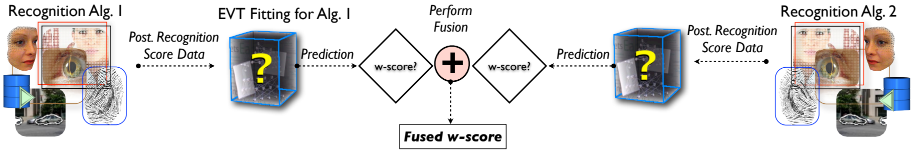

Fig. 1. An overview of the w-score normalization process. Recognition scores are produced by an algorithm for the given input. An Extreme Value Theory statistical model (Weibull) is fit to the tail of the sorted scores. The normalization of all data is done using the cumulative distribution function of the resulting Weibull distribution (hence w-scores). The w-score is an estimate of the probability of a particular score not being from the non-match distribution, and hence an ideal normalization for fusion.

probability scores based on a strong statistical theory. This is a new paradigm for recognition score normalization supporting robust recognition fusion.

We organize the rest of this paper as follows. In Section 2, we discuss the strengths and weaknesses of common recognition score normalization techniques. In Section 3, we review the post-recognition score analysis based on statistical Extreme Value Theory (pre-requisite to our new normalization technique) and in Section 4, we detail the w-score normalization technique. Finally, we present experimental results for the w-score on a series of biometric recognition algorithms and content-based image retrieval descriptors in Section 5.

2 Recognition Score Normalization

2.1 Recognition Systems

There are multiple formal ways to define what exactly a "recognition" task is. For this work, we consider the general definition of Shakhnarovich et al. [4], 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 distribution p_0 submitted to the recognition system in order to find its corresponding class label c^*. Similarly, we define *gallery* to be all the classes c^* known to the recognition system.

Many systems replace the probability in the above definition with a more generic "score," which produces the same answer when the posterior class probability of the identities is monotonic with the score function. In this case, setting the minimal threshold on a score effectively fixes δ. We call this rank-1 recognition, because the recognition is based on the largest score. One can generalize the concept of recognition, as is common in content-based image retrieval some biometrics problems and some object recognition problems, by relaxing the definition of success to having the correct answer in the top K responses. Many researchers use a pseudo-distance measure where smaller scores are better, which is trivially converted to a "larger is better" approach.

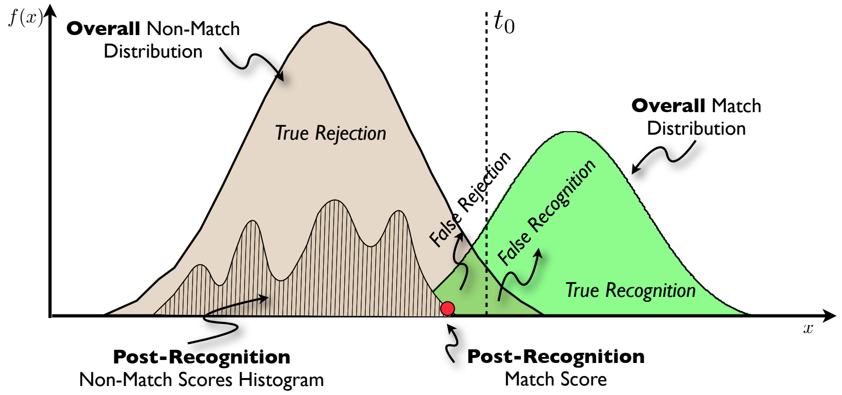

Fig. 2. The match and non-match distributions for the recognition problem. A threshold t_0 applied to the score determines the decision for recognition or rejection. Where the tails of the distributions overlap is where we find *False Rejection* and *False Recognition*. Embedded within the overall distribution is shown a particular set of post-recognition scores, with 1 match (falsely rejected by the threshold t_0) and many non-match samples.

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 Figure 2) . 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 the example yield a False Rejection for the t_0 shown.

2.2 Normalization Techniques

Traditional normalization techniques change the location and scale parameters of a score distribution. Jain et al. [5] define two types of normalizations based on the data requirements for parameter estimation. In *fixed score normalization*, which includes machine learning based approaches, the parameters used for normalization are determined *a priori* using a fixed training set. This means that the training set must accurately reflect the score distribution for each recognition algorithm – any deviation will have an impact on the recognition results. In an approach that is inline with our desire for simple parameter estimation, *adaptive score normalization* estimates parameters based on the scores at hand for a particular recognition instance. As a further consideration, a normalization technique is *robust* if it is insensitive to outliers. In this section, we briefly describe various normalization techniques, including the very popular z-score, which we use for comparison in all of our experiments in Section 5. For each example, a set of match scores $\{s_k\}, k = 1, 2, \ldots, n$ is considered for normalization.

z-scores are adaptive score normalizations that are computed in a straightforward manner. Referring to Equation 2, the normalized score is produced by subtracting the arithmetic mean μ of $\{s_k\}$ from an original score, and dividing this number by the standard deviation σ of $\{s_k\}$. This parameter estimation makes z-score normalization an adaptive score normalization, but it is possible to compute z-score normalization as a fixed score normalization if μ and σ are estimated for the overall distributions of scores produced by different recognition algorithms. z-score normalization is not robust in the traditional sense, and, as we show in this paper, is highly impacted by recognition algorithm failure.

$$s'_k = \frac{s_k - \mu}{\sigma} \tag{2}$$

tanh-estimators [6] are fixed score normalizations that are considered robust to noise, but are far more complicated to compute, compared to the adaptive z-scores. The normalized score is produced by taking the hyperbolic tangent of a z-score-like calculation. The robust nature of tanh-estimators comes from the mean and standard deviation estimates, which are computed from a genuine score distribution that is itself computed from Hampel estimators, making tanh-estimators fixed score normalizations. The Hampel estimators are based on an influence function, which makes the normalization robust to noise by reducing the influence of the scores at the tails of the distribution being considered. The tail points for three different intervals from the median score of the distribution must be defined in an *ad hoc* manner. These parameters can be difficult to determine experimentally, and if chosen incorrectly, limit the effectiveness of tanh-estimators. tanh-estimators are robust to noise, but not parameter estimation. Further, tanh-estimators have been shown to produce good results for noisy data in verification problems [5], but not recognition problems, where the underlying score distributions are different.

Other important work in score normalization has investigated advanced topics in statistical modeling including: the effect of correlation and variance on z-scores [7]; client specific normalization related to classifications made with respect to the Doddington's zoo effect (which includes failure cases) [8]; cost-sensitive performance evaluation of hardware and software failure [9]; and effects related to signal quality [10].

3 Statistical Extreme Value Theory

As we saw in Section 2.1, we 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 [4]. 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 the more general rank-K recognition). With these two definitions in mind, it is critical to note that the analysis here is done for a single probe at a time, and this assessment is not based on the overall "match/non-match" distributions, such as those in [11] and [12], which include scores over many probes. Rather it is done using a single probe producing at most one match score mixed in with a larger set of non-match scores.

We can formalize our analysis as score-based accuracy prediction for rank-K recognition, 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:

$$H_0(failure) : \forall x \in S(K), x \in \mathcal{F}(p), \tag{3}$$

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

While some researchers have formulated recognition as hypothesis testing given the individual class distributions [4], that approach presumes good models of distributions for each match/class. We cannot model the "match" distribution here effectively, as we only have *one* sample per probe, and so the only way to apply that is to assume a consistent distribution across all probes, which is questionable. That is the 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 a good non-match modeling and outlier detection. If the best score is a match it's an outlier with respect to the rest of the data.

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 [12], 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 [12] account for the bulk of the data, but have problems in the tails, and are not a good model for a particular probe.

An important observation about the problem we consider 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 are looking at the upper tail, the top n scores, there is a strong bias in the samplings that impact the tail modeling; we are interested only in the top similarity scores.

Claiming the tail of a distribution to be an extreme value problem may appear intuitive. Recent work [13] looking at verification score spaces relies on this intuition, but does not explain why extrema value theory applies to the tails of their score distributions. Just being in the tail is not sufficient to make this an extreme value problem, as one can take the top N samples from any particular distribution D, which by definition fit distribution D and not any other distribution. Just considering tails of data is not sufficient justification to invoke the extreme value theorem, just like taking a sample from a distribution does not necessarily invoke the central limit theorem.

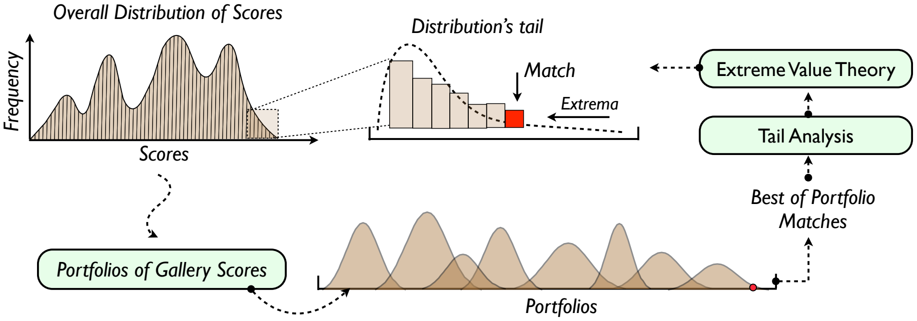

Fig. 3. Why our score analysis is an extreme value problem. One can view this problem as considering a collection of portfolios composed of sub-sets of the gallery, each of which produce scores. One portfolio contains a match-score (red), the rest are non-matching scores (brown). The best of the best 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 for the data we consider. The best score in the tail is, if a match, an outlier with respect to the EVT model of the non-match data.

We can consider the recognition problem as logically starting with a collection of portfolios, each of which 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 the subset where *all* of these scores 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 an extreme value problem. With this formalized view of recognition, we can invoke the Extreme Value Theorem: [14]:

Extreme Value Theorem 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*

$$\lim_{x \to \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x) \tag{4}$$

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

Thus, a particular portfolio is represented as the sampling $(s_1, s_2, \ldots)$, drawn from an overall distribution of scores S. Theorem 1 tells us that a large set of individual maximums 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 approximately *i.i.d.*.

The EVT is analogous to a central-limit theorem, but with minima (or maxima) over the data. 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 [15] showed that for any continuous and invertible initial distribution, only three models are needed, depending on whether you are interested in the maximum or the minimum, and also if the observations are bounded from above or below. Gumbel also proved that if a system/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

$$GEV(t) = \begin{cases} \frac{1}{\lambda} e^{-v^{-1/k}} v^{-(1/k+1)} & k \neq 0 \\ \frac{1}{\lambda} e^{-\left(x+e^{-x}\right)} & k = 0 \end{cases} \tag{5}$$

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. Equation 6 gives the CDF of a Weibull.

$$CDF(t) = 1 - e^{-\left(\frac{t}{\lambda}\right)^k} \tag{6}$$

If we presume that match scores are bounded, then the distribution of the minimum (or maximum) reduces to a Weibull (or Reversed Weibull) [16], 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 a distributional model for overall 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 Weibull distribution.*

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 extreme of their underlying portfolio, and fits a Weibull distribution to that data. Given the Weibull fit to the data, we can determine if the top score is an outlier, by considering the amount of the cumulative distribution function that is to the right of the top score.

4 Normalization via w-scores

With the necessary theory covered, we can describe the process for computing w-scores (Weibull-score, for the statistical fitting that serves as its basis) for

Algorithm 1 w-score Normalization Technique

```
Require: A collection of scores $S$, of vector length $m$, from a single recognition algorithm $j$;
Sort and retain the $n$ largest scores, $s_1, \ldots, s_n \in S$;
Fit a GEV or Weibull distribution $W_S$ to $s_2, \ldots, s_n$, skipping the hypothesized outlier;
while $k < m$ do
    $s'_k = \mathrm{CDF}(s_k, W_S)$
    $k \leftarrow k + 1$
end while
```

score normalization. The exact process for computing w-score normalization is given in Algorithm 1. The w-score re-normalizes the data based on its formal probability of being an outlier in the extreme value “non-match” model, and hence its chance of being a successful recognition. This is an adaptive score normalization; we only require the scores from a single recognition instance for a particular recognition algorithm. w-scores are very robust to noise and failure.

As w-scores are based on the fitting of the Weibull model to the non-match data of the top scores, an issue that must be addressed is the impact of any outliers on the fitting. For rank-1 fitting, where the top score is the expected match data, 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 is correct), then excluding it does not impact the fitting. If the top score was not a match, including this recognition in the fitting will bias the distribution to be broader than it should, which will produce lower probability scores for the correct match and most of the non-matches. In addition, we must address the choice of n, the tail size to be used in fitting. This tail size represents the only parameter that must be estimated for w-scores. Including too few scores might reduce accuracy, including too many items could impact assumptions of portfolio sampling. However, as we show in Section 5, even very small tail sizes (3 and 5) produce good normalization. That is consistent with work in other fields [14], where 3-5 is a very common fitting size range for Weibulls.

Once the fitting has taken place, we have all of the information necessary to complete the normalization. For every gallery class i, let score $s'_{i,j}$ be its normalized score in the collection of scores S for algorithm j. We use the CDF defined by the parameters of the fitting W_S to produce the normalized probability score $s'_{i,j}$ (we note that in Algorithm 1, the normalization process follows the sorted list of scores for a single recognition algorithm; $s'_{i,j}$ is a score-index representation for fusion). We then define w-score fusion as

$$f_i = \sum_j s'_{i,j}. \tag{7}$$

Alternatively, similar to Equation 1, one can consider the sum of only those items with a w-score (probability of success) above some given threshold δ, or could consider products or likelihood ratios of the w-scores.

Algorithm 2 w-score Error Detection For Fusion

```
Require: A collection of w-scores $S'_n$, where $n$ is the number of algorithms to fuse,
    and the collection has $m$ different score vectors for each algorithm;
Require: Algorithm FRR/FAR at current settings or ground-truth for each recognition instance;
Require: A significance threshold $\epsilon$ and an error percentage threshold $\mathcal{T}$;
while $i < m$ do
    while $j < n$ do
        $f_1 \leftarrow f_1 + s'_{i,j,1}$.
    end while
    if not a match then
        if $f_1 \geq n \times \epsilon$ then
            PossibleMatches $\leftarrow$ PossibleMatches $+1$
        end if
    end if
    $i \leftarrow i + 1$
end while
if PossibleMatches $\geq m\mathcal{T}$ then
    return System Error Detected
end if
```

The w-score fusion possesses a unique robust property, providing built-in error detection. An inverse Weibull allows us to estimate the “confidence” of particular measurement (refer to the hypothesis test of Section 3). Considering the probabilities for the top score for each algorithm, we can determine if it is highly likely that the final fused score f_1 is not a non-match; if a particular algorithm consistently fails (or the ground-truth shows it is not failing), we have evidence of a possible error, most probably some type of data misalignment. Algorithm 2 describes the process of the error detection. A count of the possible matches is kept, and if it exceeds $\mathcal{T}$ percent, we declare system error.

We have found this error detection property to be useful for indicating three possible errors: (1) the Weibull fitting is inaccurate for valid score data (due to a mis-estimated tail size) (2) invalid score data (from parsing errors) produced a CDF that returns an improbably large number of high w-scores; (3) an error is present in alignment or the ground-truth labeling (off-by-one errors due to bad pre-processing). To our knowledge, no other fusion technique has this property.

5 Experimental Results

In this section, we present experimental results for our w-score method on a series of biometric recognition algorithms and content-based image retrieval descriptors. We compare the w-score approach to the well known *z-score* normalization. *z-score* normalization remains one of the most popular normalization techniques out there (a search for “z-score” using Google scholar returns 102,000 scholarly works), because of its theoretical performance on Gaussian data, and its straightforward parameter estimation. Fixed score normalizations such as tanh-estimators or machine learning based approaches are not considered for the reasons given in Section 2.2.

5.1 Biometric Recognition

For our first set of experiments, we tested a series of biometric recognition algorithms from the NIST BSSR1[17] biometric score set. The data set consists of scores from 2 face recognition algorithms (labeled C & G) and 1 fingerprint recognition algorithm applied to two different fingers (labeled LI & RI). BSSR1's multibiometric subset contains 517 score sets for each of the algorithms, with common subjects between each set. BSSR1 also contains individual score subsets for all algorithms, where the scores do not have common subjects between them. Out of this individual score set data, we created a "Chimera" data set with 3000 score sets and consistent labeling across all algorithms. This was done to address the limited nature of the true multibiometric set, where fusion pushes the recognition rate close to 100% for even weak normalizations.

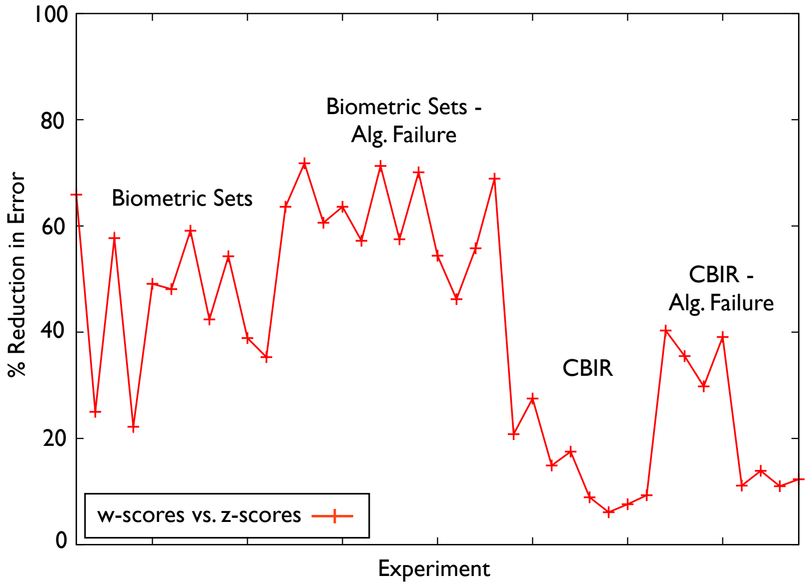

Fig. 4. A graphical summary of all of the results presented in this paper. In all cases, w-scores reduce the margin of error after fusion, when compared to z-scores (baseline), for a variety of biometric recognition algorithms and CBIR descriptors.

We performed two different types of experiments on this data. All results are presented as a percentage of error reduction (improvement) compared to z-scores, the most popular type of adaptive score normalization, calculated as

$$\%\text{reduction} = (\%e_{\text{z}} - \%e_{\text{w}})/\%e_{\text{z}} \quad (8)$$

where $\%e_{\mathrm{z}}$ is the percentage of incorrect rank-1 results for z-score fusion, and $\%e_{\mathrm{w}}$ is the percentage of incorrect rank-1 results for w-score fusion.

For the first experiment, we fused a variety of face and fingerprint recognition algorithms. We note that in normalization and fusion, performance varies as a function of the data considered. Thus, we only considered the scores equal to a percentage of the total number of classes, expressed as $\%c^*$. This threshold is independent of the Weibull fitting, and is applied to both the w-score and z-score. While we show results for experiments with a consistent percentage of classes for w-scores and z-scores, we note that in our broader experimentation, we were always able to achieve better performance than z-scores when choosing the correct tail size for fitting, and fusing scores within the tail used for fitting. The tail size used for fitting for all biometrics experiments in this paper is 5.

Algorithm	**Improve**	$\%c^*$
C & LI	65.9%	2.0%
C & RI	25.0%	2.0%
G & LI	57.7%	2.0%
G & RI	22.2%	2.0%
Chimera C & G	49.1%	0.2%
Chimera C & LI	48.1%	0.2%

Algorithm	**Improve**	$\%c^*$
Chimera C & RI	59.1%	0.2%
Chimera G & LI	42.4%	0.2%
Chimera G & RI	54.3%	0.2%
Chimera C & G & LI	38.9%	0.2%
Chimera C & G & RI	35.3%	0.2%

Table 1. Rank-1 fusion results, expressed as the percentage of error reduction compared to z-scores, for the BSSR1 multibiometric and the BSSR1 "Chimera" data sets.

The second experiment tests fusion behavior in a failure scenario, where rank-1 recognition for at least one of the algorithms is 0%. For biometrics, this may be thought of as an "impostor" test, where a subject is trying to actively defeat the recognition system (consider the possibility of a facial disguise that causes a face algorithm to fail, but has no effect on a fingerprint recognition algorithm). Results for the BSSR1 multibiometric set and the BSSR1 Chimera set are given in Tables 1 & 2. w-scores have a clear advantage over z-scores for regular fusion, and a significant advantage in cases where a recognition algorithm is failing.

Algorithm	**Improve**	$\%c^*$
*C & LI	63.6%	2.0%
*C & RI	71.8%	2.0%
*G & LI	60.6%	2.0%
*G & RI	63.6%	2.0%
Chimera *C & LI	57.2%	0.3%
Chimera *C & RI	71.3%	0.3%

Algorithm	**Improve**	$\%c^*$
Chimera *G & LI	57.5%	0.3%
Chimera *G & RI	70.1%	0.3%
Chimera LI & *RI	54.4%	0.3%
Chimera RI & *LI	46.2%	0.3%
Chimera *C & *G & LI	55.8%	0.3%
Chimera *C & *G & RI	68.9%	0.3%

Table 2. Rank-1 fusion results, expressed as the percentage of error reduction compared to z-scores, for the BSSR1 multibiometric and the BSSR1 "Chimera" data sets, fusing with failing algorithms (marked with *). Note the significant reduction in error.

5.2 Content Based Image Retrieval

To show the broader applicability of w-score normalization, we also tested a series of simple CBIR descriptors [18]. Data from the Corel "Relevants" set [19], containing 50 unique classes, and the INRIA "Holidays" set [20], containing 500 unique classes. Using a variety of descriptors, we generated 1624 score sets for Corel Relevants and 1491 score sets for INRIA Holidays. In total, we tested 47 different combinations of descriptors across all experiments, but due to space constraints, we only show four different representative combinations. The experiments are identical to those of the biometric sets in Section 5.1. Results for the Corel Relevants set and the INRIA Holidays set are given in Tables 3 & 4. We note that in all of our fusion experiments with CBIR descriptors, w-scores outperformed z-scores when the appropriate tail size was chosen for Weibull fitting, which is consistent with our biometric results. The tail sized used for fitting for all CBIR experiments is 3.

CBIR Algorithm	**Improve**	$\%c^*$
Relevants csd & gch	20.8%	6.0%
Relevants csd & jac	27.5%	6.0%
Relevants cw_{hsv} & cw_{luv}	14.9%	6.0%
Relevants cw_{hsv} & jac	17.5%	6.0%

CBIR Algorithm	**Improve**	$\%c^*$
Holidays csd & gch	8.9%	0.6%
Holidays csd & jac	6.1%	0.6%
Holidays cw_{hsv} & cw_{luv}	7.6%	0.6%
Holidays cw_{hsv} & jac	9.3%	0.6%

Table 3. Rank-1 CBIR fusion results, expressed as the percentage of error reduction compared to z-scores, for the Corel Relevants and INRIA Holidays data sets. We note that fusion performance here is relative to data set difficulty.

CBIR Descriptor	**Improve**	$\%c^*$
Relevants *csd & gch	40.3%	6.0%
Relevants csd & *jac	35.5%	6.0%
Relevantscw_{hsv} & $*cw_{luv}$	29.8%	6.0%
Relevants cw_{hsv} & jac	39.1%	6.0%

CBIR Descriptor	**Improve**	$\%c^*$
Holidays *csd & gch	11.1%	0.6%
Holidays csd & *jac	13.9%	0.6%
Holidays cw_{hsv} & $*cw_{luv}$	11.0%	0.6%
Holidays $*cw_{hsv}$ & jac	12.3%	0.6%

Table 4. Rank-1 CBIR fusion results, expressed as the percentage of error reduction compared to z-scores, for the Corel Relevants and INRIA Holidays data sets, fusing with failing algorithms (marked with *). Note the significant reduction in error for this experiment, which is consistent with the biometric results presented in Table 2

6 Conclusion

In this paper, we have introduced a theory of post-recognition score analysis based on statistical Extreme Value Theory, and used this theory to develop our new w-score adaptive score normalization. Through our analysis, we showed that no matter what model best fits each non-match distribution, with enough samples and enough classes, the sampling of the top-n scores always results in a Weibull distribution. With this knowledge, we developed a method that takes the tail of these scores, which are likely to have been sampled from the extreme of their underlying sub-sets from the gallery, and fits a Weibull distribution to that data; the CDF of the resulting distribution allows us to normalize the entire score sequence. In essence, the w-score normalizes scores to a probability score reflecting the confidence of the score not being a non-match. Results on a wide range of biometric and CBIR data show that the w-score is superior to the z-score, the most popular type of adaptive score normalization, especially when one or more recognition algorithms fail or when there are impostor scores.

References

1. Huber, P.: Robust Statistics. Wiley, New York (1981)
2. Li, W., Gao, X., Boult, T.: Predicting Biometric System Failure. In: CIHSPS. (2005)
3. Wang, P., Ji, Q., Wayman, J.: Modeling and Predicting Face Recognition System Performance Based on Analysis of Similarity Scores. IEEE TPAMI **29** (2007) 665–670
4. Shakhnarovich, G., Fisher, J., Darrell, T.: Face Recognition From Long-term Observations. In: ECCV. (2002) 851–868
5. Jain, A., Nandakumar, K., Ross, A.: Score Normalization in Multimodal Biometric Systems. Pattern Recognition **38** (2005) 2270–2285
6. Hampel, F., Rousseeuw, P., Ronchetti, E., Stahel, W.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
7. Poh, N., Bengio, S.: How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks? IEEE. TSP **53** (2005) 4384–4396
8. Poh, N., Kittler, J.: Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems. IEEE. TASLP **16** (2008) 594–606
9. Poh, N., Bourlai, N., Kittler, J.: Benchmarking Quality-Dependent and Cost-Sensitive Score-Level Multimodal Biometric Fusion Algorithms. IEEE. TIFS **4** (2009) 849–866
10. Poh, N., Bourlai, T., Kittler, J.: A Multimodal Biometric Test Bed for Quality-dependent, Cost-sensitive and Client-specific Score-level Fusion Algorithms. Pattern Recognition **43** (2010) 1094–1105
11. Shi, Z., Kiefer, F., Schneider, J., Govindaraju, V.: Modeling Biometric Systems Using the General Pareto Distribution (GPD). In: SPIE. Volume 6944. (2008)
12. Grother, P., Phillips, P.: Models of Large Population Recognition Performance. In: IEEE CVPR. (2004) 68–75
13. Broadwater, J., Chellappa, R.: Adaptive Threshold Estimation Via Extreme Value Theory. IEEE TSP (2009) To appear.
14. Kotz, S., Nadarajah, S.: Extreme Value Distributions: Theory and Applications. 1 edn. World Scientific Publishing Co. (2001)
15. Gumbel, E.: Statistical Theory of Extreme Values and Some Practical Applications. Number National Bureau of Standards Applied Mathematics in 33. U.S. GPO, Washington, D.C. (1954)
16. NIST: NIST/SEMATECH Handbook of Statistical Methods. 33. U.S. GPO (2008)
17. NIST: Biometric Scores Set (2004) www.itl.nist.gov/iad/894.03/biometricscores.
18. Datta, R., Joshi, D., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM CSUR **40** (2008) 1–77
19. Stehling, R., Nascimento, M., Falcão, A.: A compact and efficient image retrieval approach based on border/interior pixel classification. In: CIKM. (2002) 102–109
20. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometry consistency for large scale image search. In: ECCV. (2008)