Robust Fusion: Extreme Value Theory for Recognition Score Normalization

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

[1] Univ. of Colorado at Colorado Springs & Securics, Inc , US | [2]Univ. of Campinas, Brazil | [3]National Inst. of Standards and Technology, US

Motivation

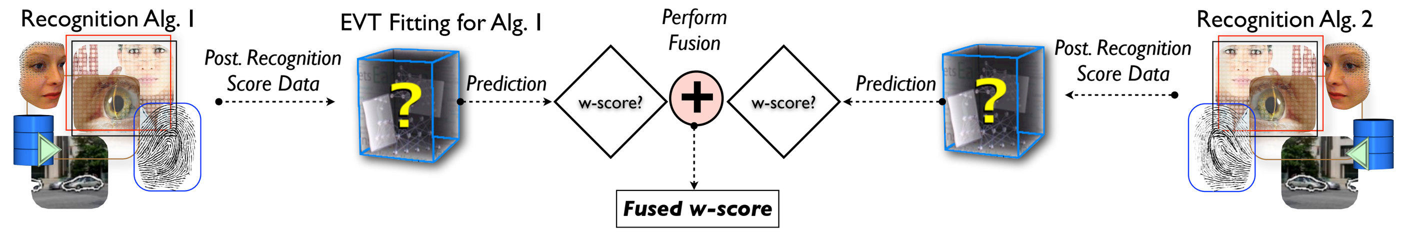
We introduce the **w-score**, a new technique for robust score normalization that does not assume a match or non-match distribution, but instead relies on the observation that the scores of a recognition system's non-match scores follow the statistical Extreme Value Theory.

Our goal is to achieve **robust fusion**: a fusion process that is insensitive to errors in its distributional assumptions on the data, has simple parameter estimation, and a high input failure tolerance.

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

1. The varying nature of the underlying distribution of scores across different recognition algorithms often leads to inconsistent normalization results.
2. Complications arise when one or more sensors or recognition algorithms considered for fusion fail or are deceived.

The w-score is designed to address both of these problems



An overview of the w-score process. Our robust fusion is inspired by post-recognition score analysis[1,2].

Recognition Systems

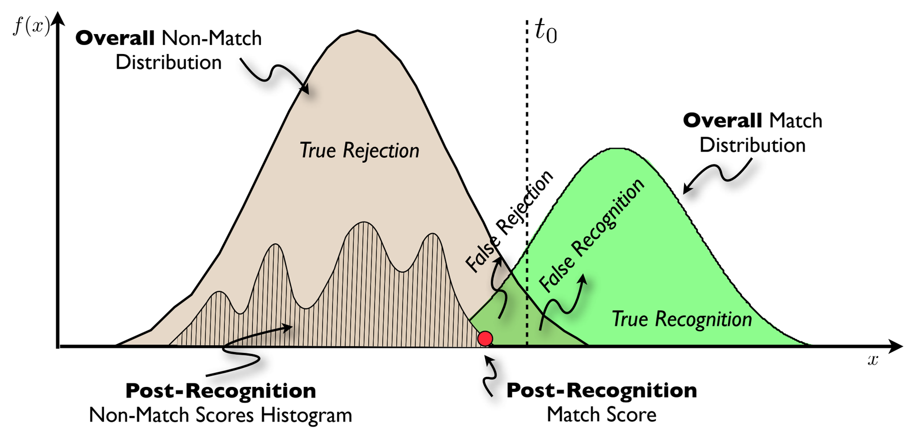
The task of a recognition system[3] is to find the class label c^*, where p_k is an underlying probability rule and p_0 is the input distribution, satisfying

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

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.

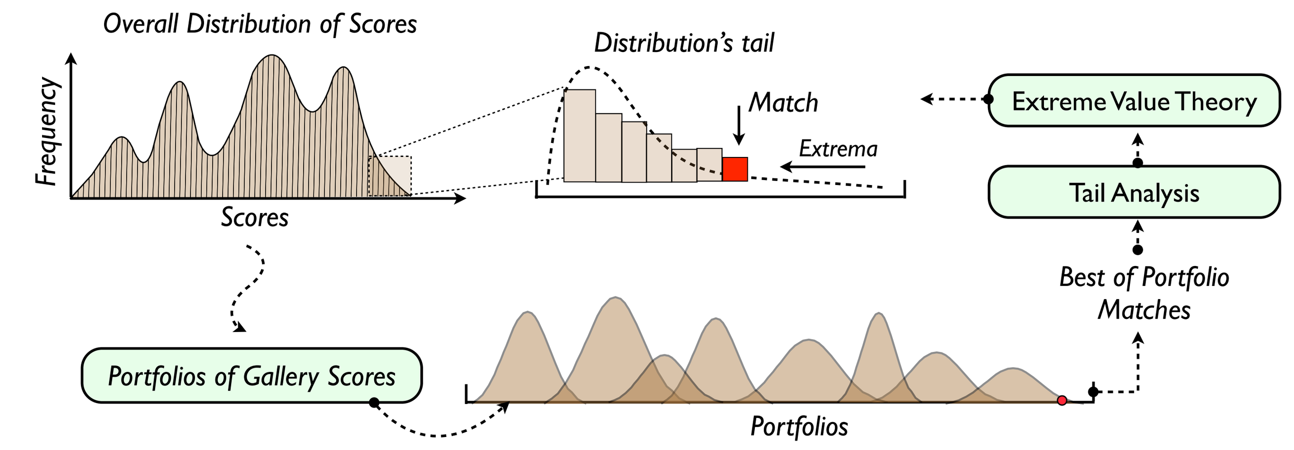
probe is defined as the input image distribution p_0 submitted to the recognition system in order to find its corresponding class label c^*.

gallery is defined to be all the classes c^* known to the recognition system.



Where the tails of the match and non-match distributions overlap is where we find *False Rejection* and *False Recognition*.

Statistical Extreme Value Theory



Consider a collection of portfolios composed of sub-sets of gallery 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 score distribution. The best score in the tail is, if a match, an outlier with respect to the Extreme Value Theory[4] model of the non-match data.

Extreme Value Theorem 1 *Let* $(s_1, s_2, \dots)$ *be a sequence of i.i.d. samples. Let* $M_n = \max\{s_1, \dots, s_n\}$. If a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and

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

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

Theorem 1 tells us that a large set of individual maximums M_n from the portfolios must converge to an extreme value distribution. These maxima are most accurately modeled by the Weibull Distribution (Extreme Value Type III Distribution).

W-Score Normalization

Algorithm 1: w-score normalization

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

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.

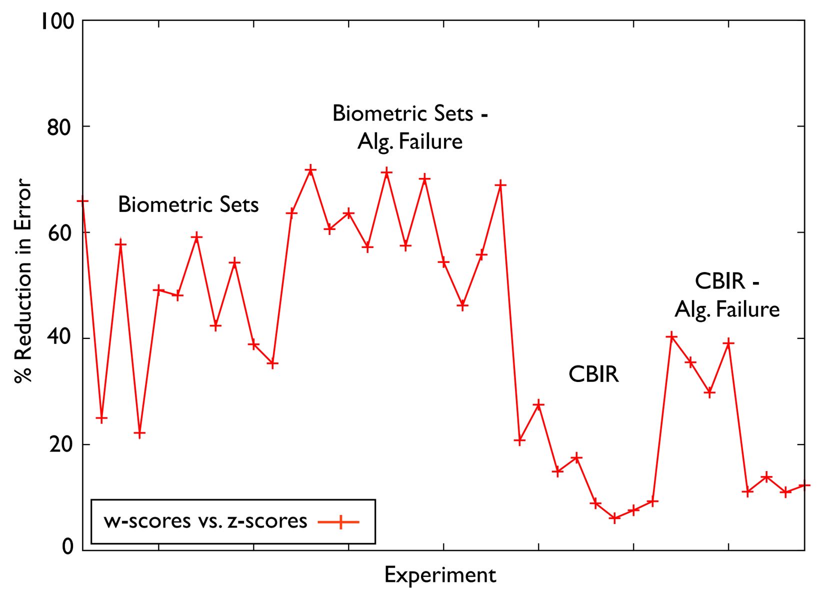
We use the CDF defined by the parameters of the fitting to produce the normalized probability score. The tail size used for fitting is the only parameter that must be estimated empirically.

Algorithm 2: w-score error detection for fusion

Require: A collection of w-scores S'_n, with m different score vectors per algorithm, and n different algorithms;
Require: Algorithm FRR/FAR at current settings or ground-truth;
Require: A significance threshold ε and an error percentage threshold T;
while $i < m$ **do**
  while $j < n$ **do**
    $f_i \leftarrow f_i + s'_{i,j,1}$ # f_i is the fused score
  end while
  if not a match **then**
    if $f_i \geq n \times \varepsilon$ **then**
      $\mathrm{PossibleMatches} \leftarrow \mathrm{PossibleMatches} + 1$
    end if
  end if
  $i \leftarrow i + 1$
end while
if $\mathrm{PossibleMatches} \geq mT$ **then**
  return System Error Detected
end if

Detection for three possible errors: (1) inaccurate Weibull fitting, (2) invalid score data produced CDF that returns too many high w-scores (3) data alignment error

Experimental Results



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.

Biometric Recognition: NIST BSSR1 biometric score set; 2 face recognition algorithms (labeled C & G) and 1 fingerprint recognition algorithm applied to two different fingers (labeled LI & RI); 517 scores across common subjects in true multibiometric set, 3000 scores across common subjects in our Chimera set.

Content Based Image Retrieval: Corel "Relevants" set, containing 50 unique classes, and the INRIA "Holidays" set, containing 500 unique classes; using a variety of descriptors, we generated 1624 score sets for Corel Relevants and 1491 score sets for INRIA Holidays.

All sum rule results are presented as a percentage of error reduction (improvement) compared to z-scores[5], the most popular type of adaptive score normalization, calculated as: $\%\text{reduction} = (\%e_z - \%e_W) / \%e_z$

w-scores are most useful when some of the data considered for fusion is "failing".

Algorithms	Improvement	$\%c^*$
!C & LI	63.6%	2.0%
!C & RI	71.8%	2.0%
!G & LI	60.6%	2.0%
!C & RI	63.6%	2.0%
Chimera !C & LI	57.2%	0.3%
Chimera !C & RI	71.3%	0.3%

Algorithms	Improvement	$\%c^*$
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%

Rank-1 fusion results, compared to z-scores, for the BSSR1 multibiometric and the BSSR1 "Chimera" data sets, fusing with failing algorithms (marked with !).

CBIR Descriptors	Improvement	$\%c^*$
Relevants !csd & gch	40.3%	6.0%
Relevants csd & !jac	35.5%	6.0%
Relevants cw_{hsv} & !cw_{luv}	29.8%	6.0%
Relevants !cw_{hsv} & cw_{luv}	39.1%	6.0%

CBIR Descriptors	Improvement	$\%c^*$
Relevants !csd & gch	11.1%	0.6%
Relevants csd & !jac	13.9%	0.6%
Relevants cw_{hsv} & !cw_{luv}	11.0%	0.6%
Relevants !cw_{hsv} & cw_{luv}	12.3%	0.6%

Rank-1 CBIR fusion results, compared to z-scores, for the Corel Relevants and INRIA Holidays data sets, fusing with failing algorithms.

$\%c^*$ represents the tail size equal to a percentage of the total number of classes

References

1. Li, W., Gao, X., Boult, T.: *Predicting Biometric System Failure.* In: CIHSPS. (2005)
2. Scheirer, W. and Boult, T.: *A Fusion Based Approach to Enhancing Multi-Modal Biometric Recognition System Failure Prediction and Overall Performance.* In: BTAS. (2008)
3. Shakhnarovich, G., Fisher, J., Darrell, T.: *Face Recognition From Long-term Observations.* In: ECCV. (2002) 851–868
4. 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)
5. Jain, A., Nandakumar, K., Ross, A.: *Score Normalization in Multimodal Biometric Systems.* Pattern Recognition **38** (2005) 2270–2285

Acknowledgements

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 Grants 2008/08681-9 and 2010/09825-4. We also extend our thanks to J. Ross Beveridge for providing valuable feedback on an earlier draft of this work.