Statistical Methods for Open Set Recognition

Walter J. Scheirer and Terrance E. Boult

Part 2: Statistical Extreme Value Theory for Visual Recognition

The Statistical Extreme Value Theory (EVT)

Why EVT for visual recognition problems?

- Powerful explanatory theory (Scheirer et al. T-PAMI 2011)
- Effective tool for statistical modeling of decision boundaries (Broadwater et al. IEEE T. Signal Processing 2010, Fragoso and Turk CVPR 2013)
  - Calibration models (Scheirer et al. ECCV 2010)

The Extreme Value Theorem

Let $(s_1, s_2, \ldots, s_n)$ 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

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

then if F is a non-degenerate distribution function, it belongs to one of three extreme value distributions[1].

The i.i.d. constraint can be relaxed to a weaker assumption of exchangeable random variables[2].

1. S. Kotz and S. Nadarajah, Extreme Value Distributions: Theory and Applications, 1st ed. World Scientific Publishing Co., 2001.

2. S. Berman, "Limiting Distribution of the Maximum Term in Sequences of Dependent Random Variables," Ann. Math. Stat., vol. 33, no. 3, pp. 894-908, 1962.

The Weibull Distribution

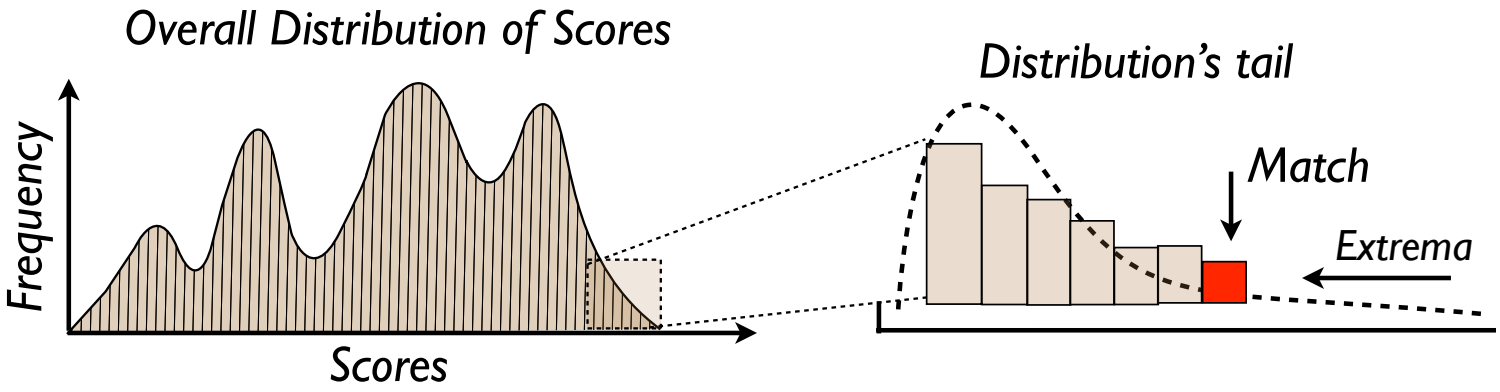
The sampling of the top-n scores always results in an EVT distribution, and is ***Weibull*** if the data are bounded[1].

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Choice of this distribution is not dependent on the model that best fits the entire non-match distribution.

1. NIST/SEMATECH e-Handbook of Statistical Methods, ser. 33. U.S. GPO, 2008

Fitting an EVT Distribution



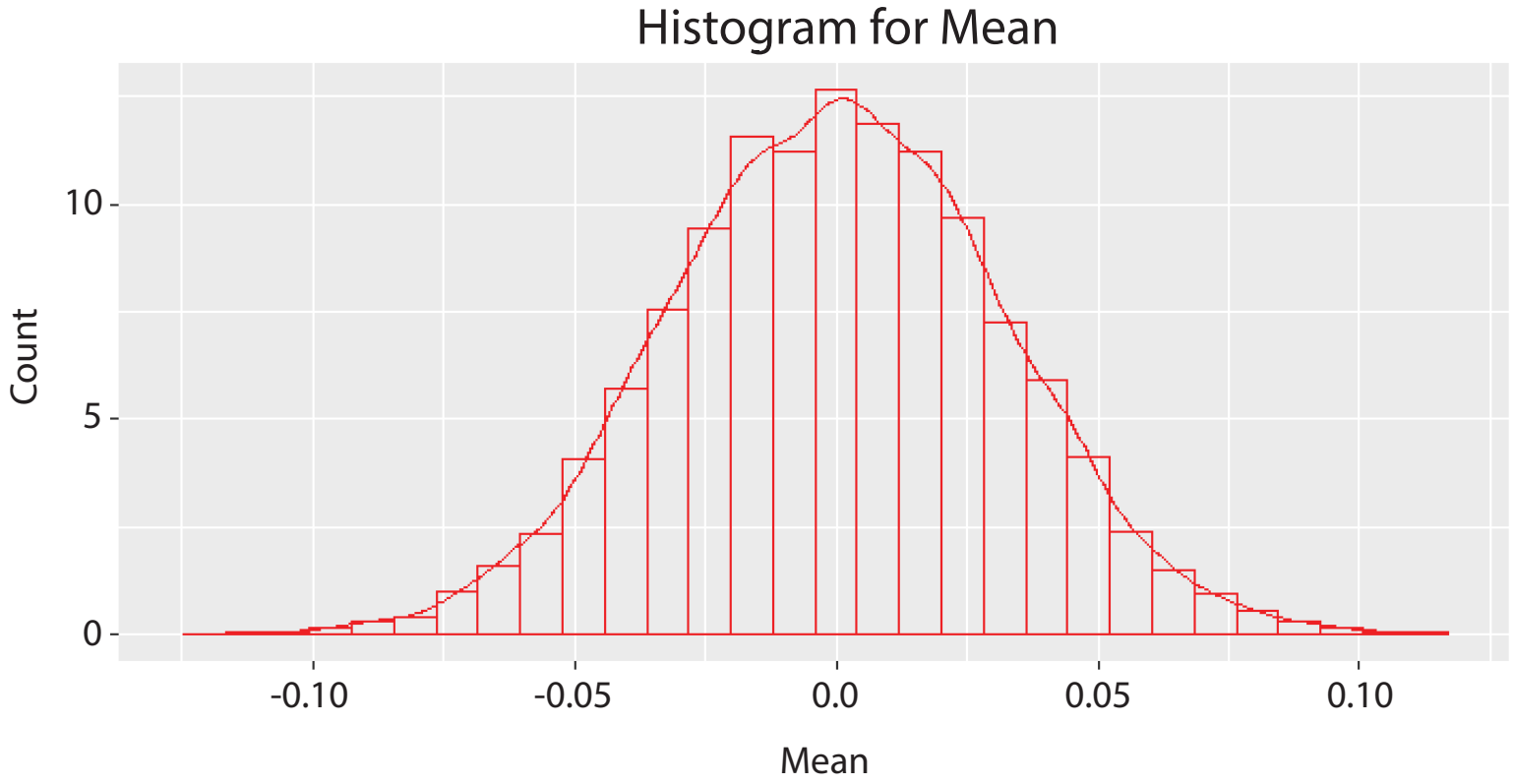
- EVT applies regardless of the overall distribution
- Sampling the extrema in the tail of an overall distribution always results in an EVT distribution

Is there a difference between central tendency modeling and EVT?

- Sample set of 1,000 values from a standard normal distribution
- Compute means over 10,000 trials

What does the histogram look like?

Bell curve

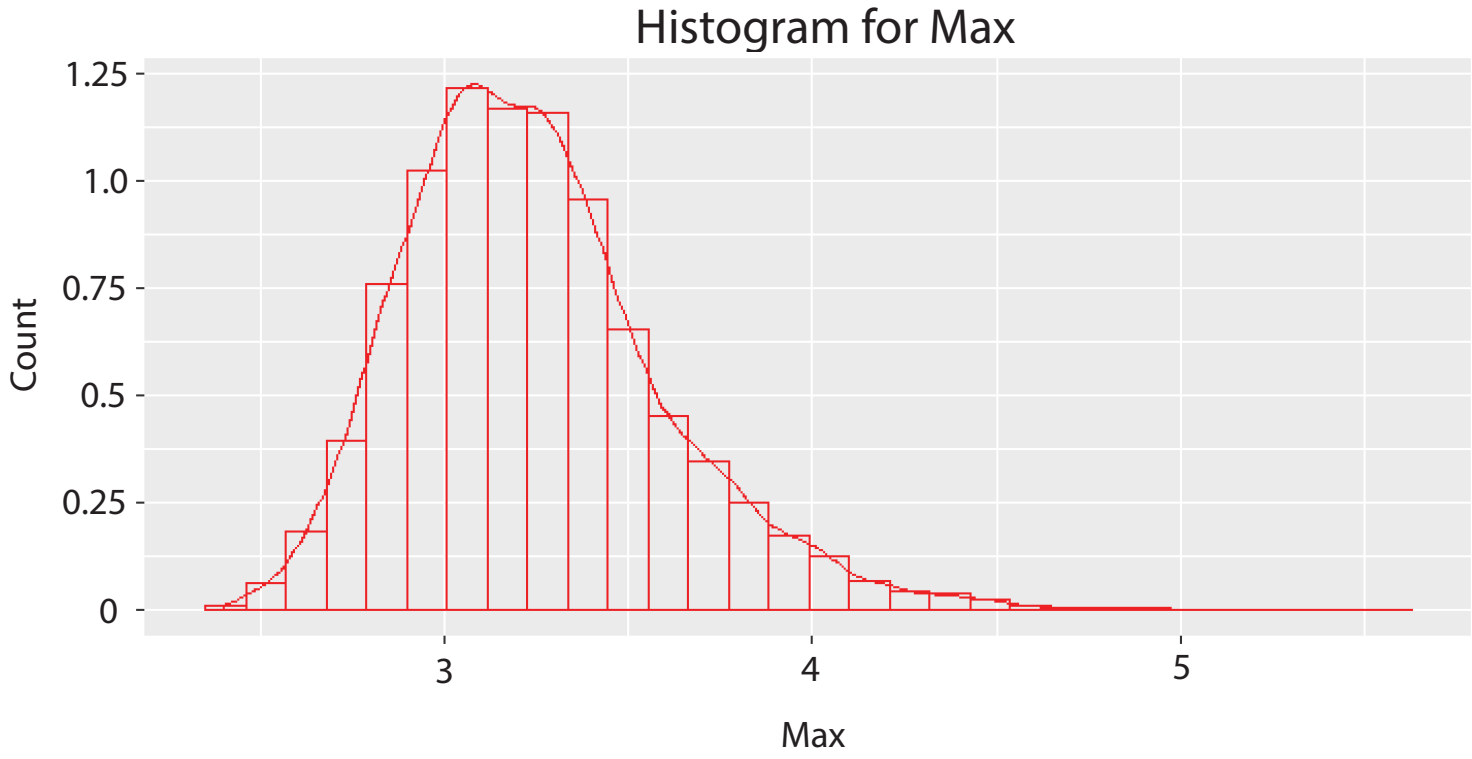


What if we're interested in extrema points instead?

- Sample set of 1,000 values from a standard normal distribution
- Retain the *maximums* over 10,000 trials

What does the histogram look like?

The peak is now at 3.2, and there is noticeable skew



Probability estimation (R): central tendency

Means from the 10,000 trials

```
library("MASS")
fitdistr(bufferMean, "normal")
#     mean           sd
# -0.0001344194 0.0313828480
# (0.0003138285) (0.0002219102)
```

Improbable outcome

```
pnorm(0, -0.0001344194, 0.0313828480, FALSE)
# [1] 0.4982913
pnorm(1, -0.0001344194, 0.0313828480, FALSE)
# [1] 3.611041e-223
pnorm(2, -0.0001344194, 0.0313828480, FALSE)
# [1] 0
```

Probability estimation (R): EVT

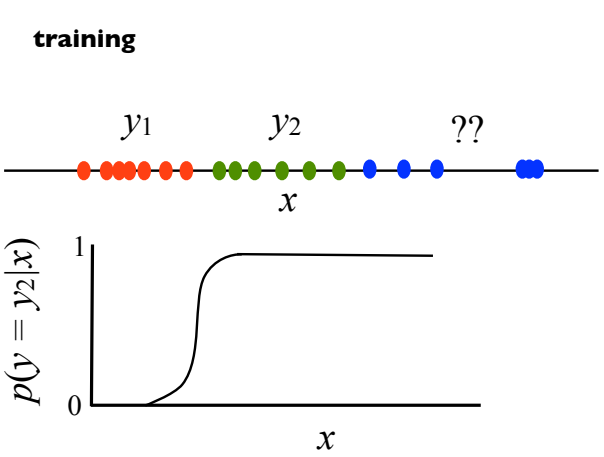
Maximums from the 10,000 trials

```
library(SpatialExtremes)
gevmle(bufferMax)
    loc        scale        shape
3.08305916 0.29802546 -0.07158273
```

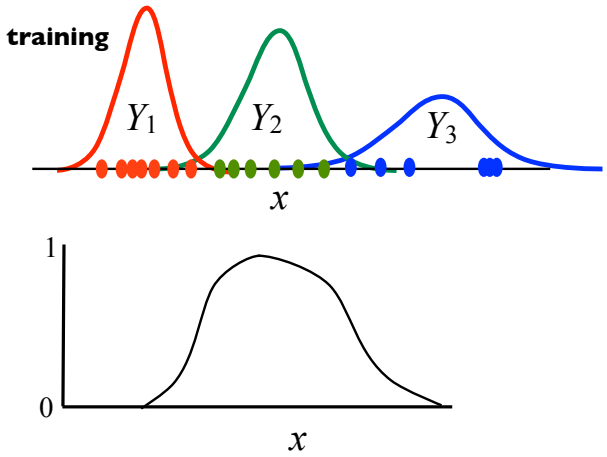
Probable outcome

```
pgev(0, 3.08305916, 0.29802546, -0.07158273, FALSE)
#[1] 1
pgev(1, 3.08305916, 0.29802546, -0.07158273, FALSE)
#[1] 1
pgev(2, 3.08305916, 0.29802546, -0.07158273, FALSE)
#[1] 1
pgev(3, 3.08305916, 0.29802546, -0.07158273, FALSE)
#[1] 0.7322732
pgev(4, 3.08305916, 0.29802546, -0.07158273, FALSE)
#[1] 0.03047941
```

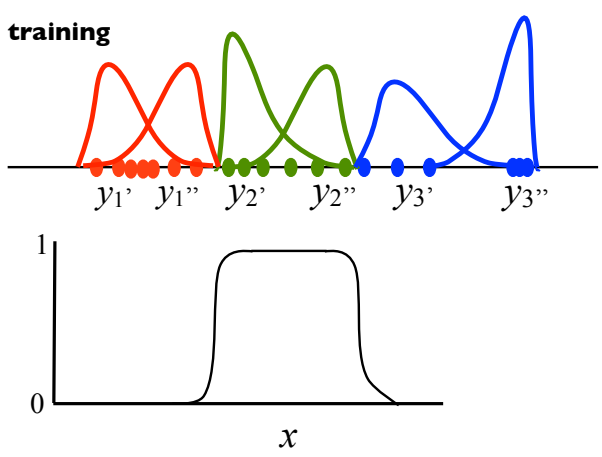
A good alternative to central tendency modeling



(a) Binary Discriminative Model

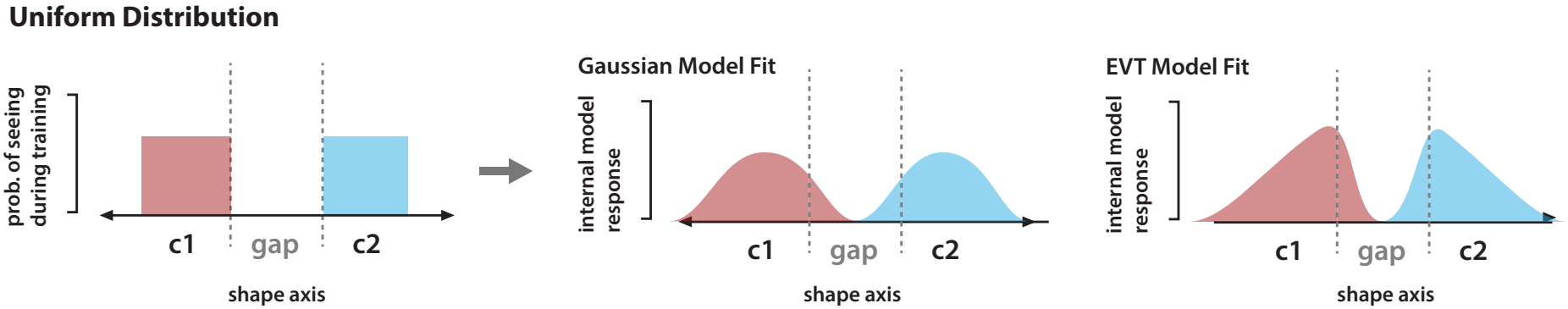


(b) Per class Gaussian Model + Bayesian decision



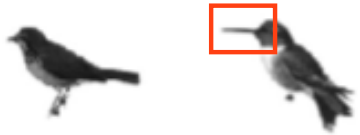
(c) EVT Fit for the min and max tail of each class + Bayesian decision

Example two-category discrimination task along a parametric stimulus axis



Extrema as visual features

Tanaka et al. - Atypicality

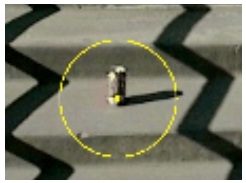


Typical Atypical

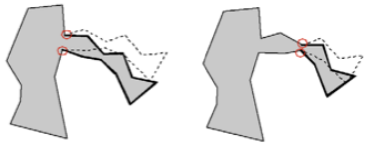
Leopold et al. - Caricaturization

Tanaka and Farah - Visual Attributes

Itti et al. - Visual Saliency



Barenholtz and Tarr - Part Boundaries



E. Barenholtz and M. Tarr. Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects. Acta Psychologica, 128:331– 338, May 2008.

J.W. Tanaka and M.J. Farah. Parts and wholes in face recognition. Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 46:225–245, 1993.

D. Leopold, I. Bondar, and M. Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature, 442:572–575, August 2006.

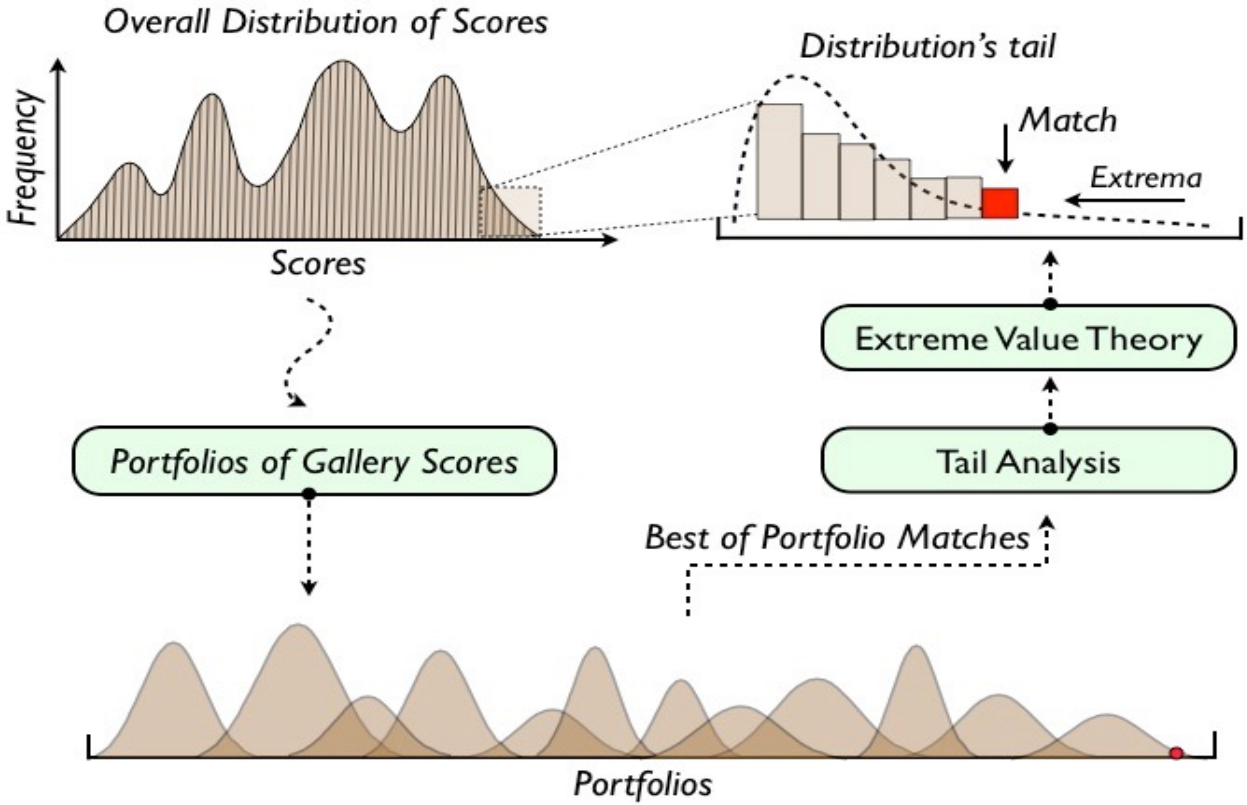
L. Itti and C. Koch. Computational Modeling of Visual Attention. Nature Reviews Neuroscience, 2(3):194–203, February 2001.

J.W. Tanaka and O. Corneille. Typicality effects in face and object perception: Further evidence for the attractor field model. Perception & Psychophysics, 69(4):619–627, May 2007.

How does EVT apply to computer vision?

Meta-Recognition Theory

Meta-recognition is **recognizing when a recognition system is working or failing**. It is important for threshold selection, failure prediction and improving fusion.



W. J. Scheirer, A. Rocha, R. J. Micheals, T. E. Boult, "Meta-Recognition, the Theory and Practice of Recognition Score Analysis," vol. 33, no. 8, 2011

Failure Prediction

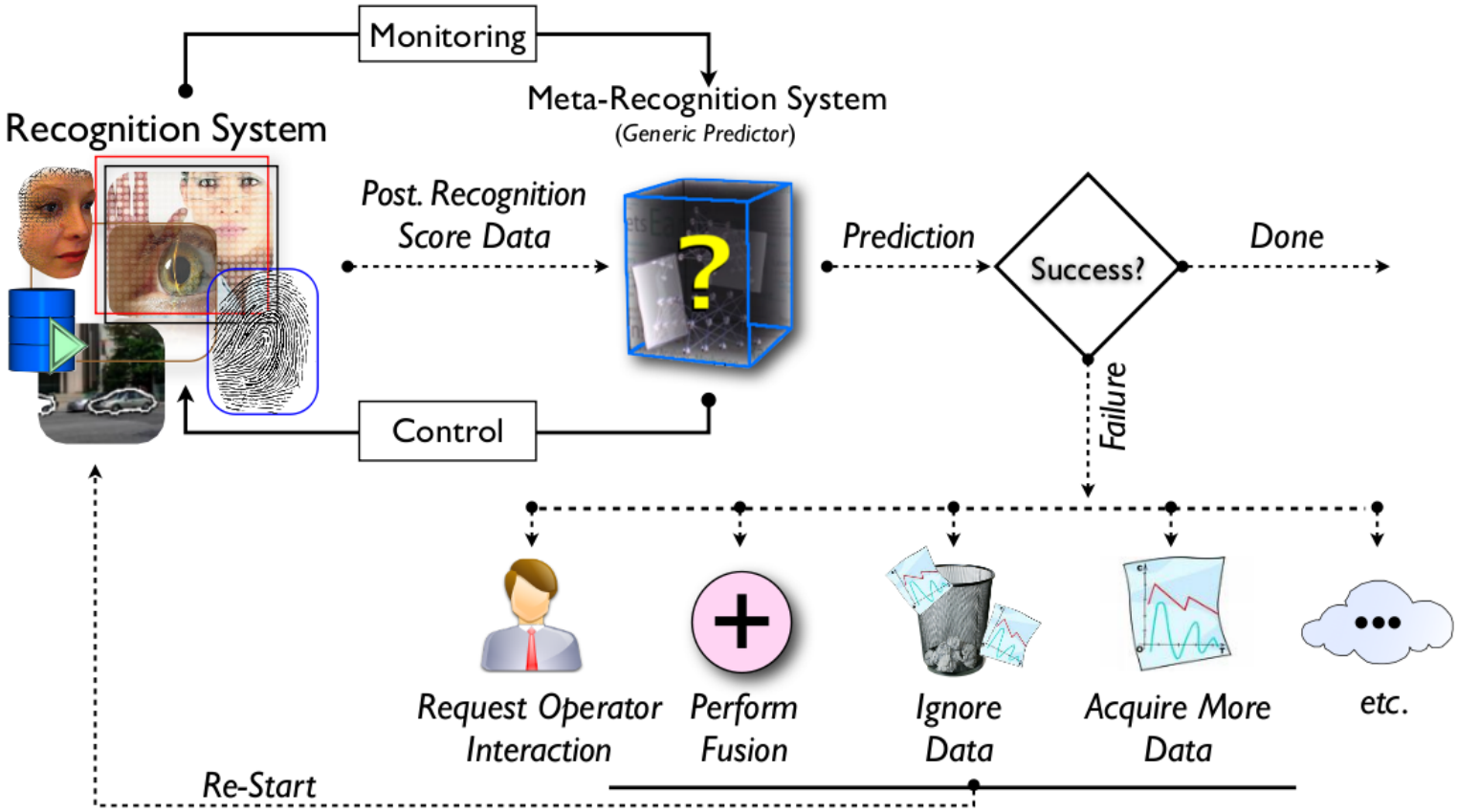
Can we recognize, in some automated fashion, if a recognition system result is a success or a failure?

If so, can we quantify the probability of success or failure?

Success or Failure?

Success or Failure?

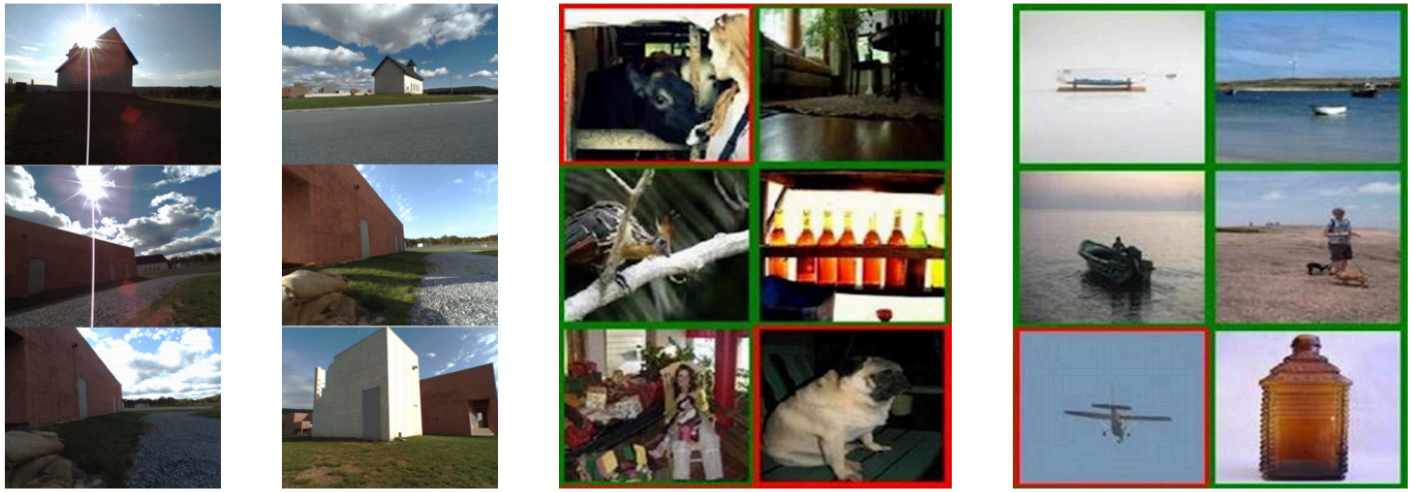
Meta-Recognition as failure prediction



Predicting Failures of Vision Systems

Zhang et al. CVPR 2014

Learn conditions that cause a target algorithm to fail



(a) Predicted Bad

(b) Predicted Good

(c) Predicted Bad

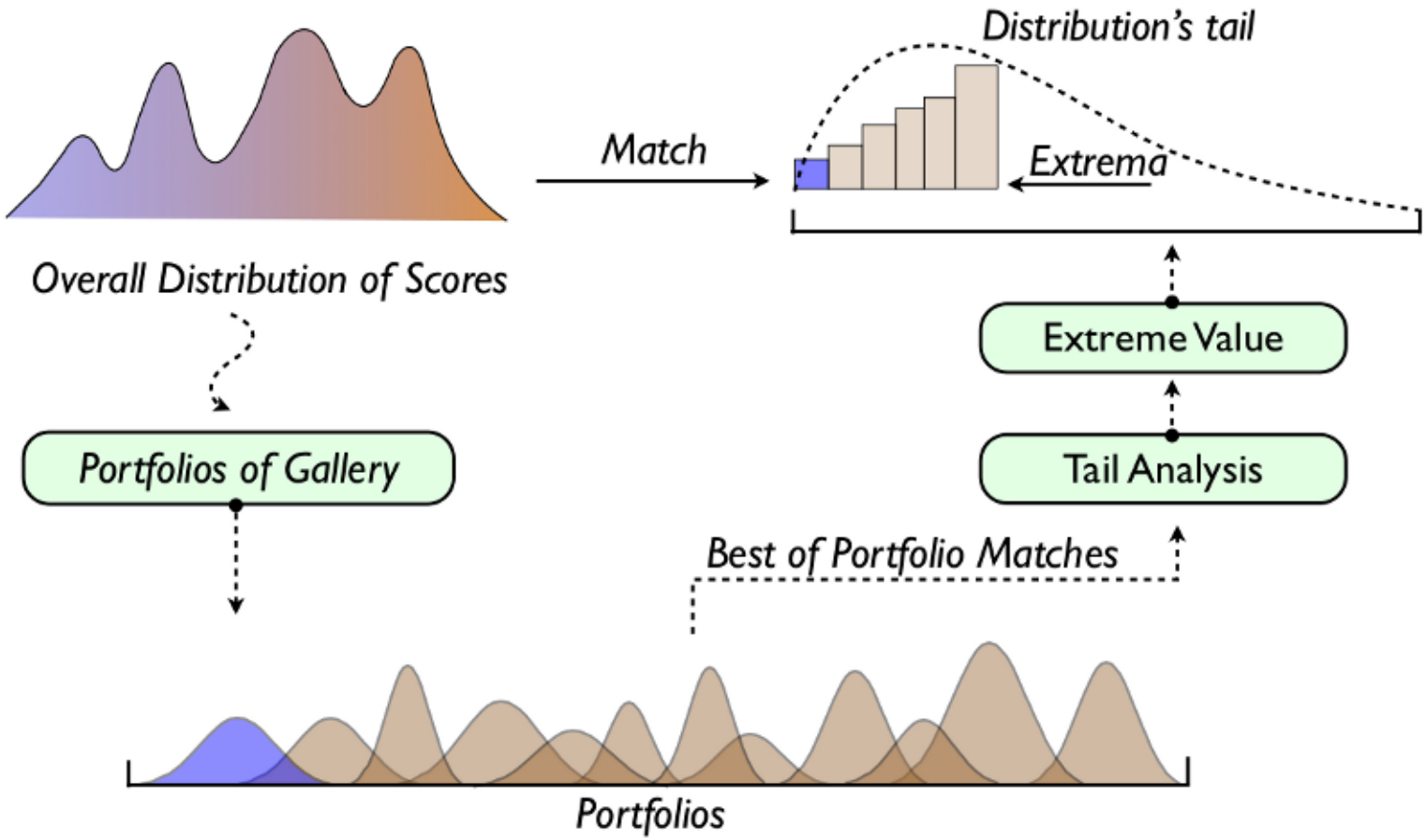
(d) Predicted Good

P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting Failures of Vision Systems," CVPR 2014

Statistical EVT Failure Prediction

- Get scores, sort and take top N
- Fit an extreme value distribution to get model of non-match distribution, exclude top score
- Determine if top score is outlier from distribution, If so predict success. Else predict failure
- Detect outlier using fraction CDF below the potential outlier. 99.99999% is a good test!

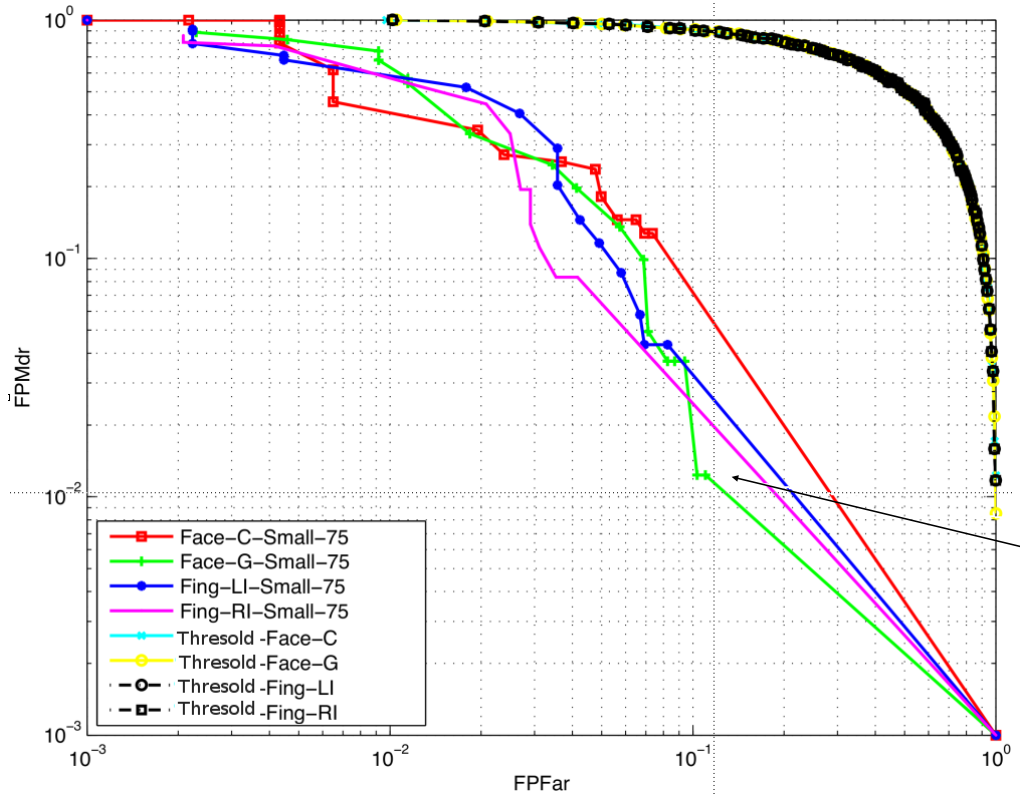
Statistical EVT Failure Prediction



EVT-based failure prediction

- Using meta-recognition, evaluated 12 algorithms across 4 problems. Always significantly better than simple thresholds on score.
  - ✓Face Biometrics
  - ✓Fingerprint Biometrics
  - ✓Multi-biometric fusion
  - ✓SIFT + earth-mover distance based object recognition
  - ✓Content-based image retrieval (4 algorithms)
- Led to new fusion algorithm, better than traditional algorithms on all datasets considered.

Example prediction accuracy

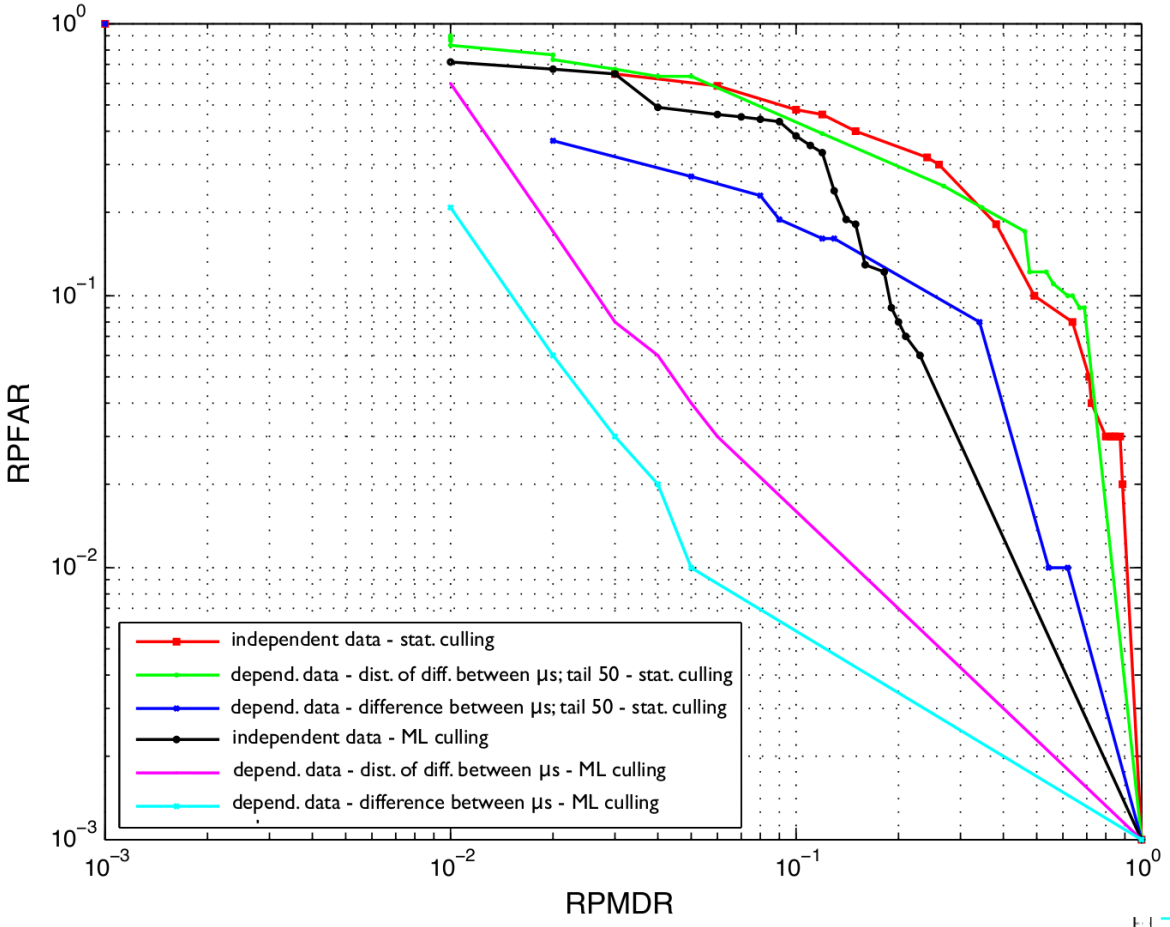


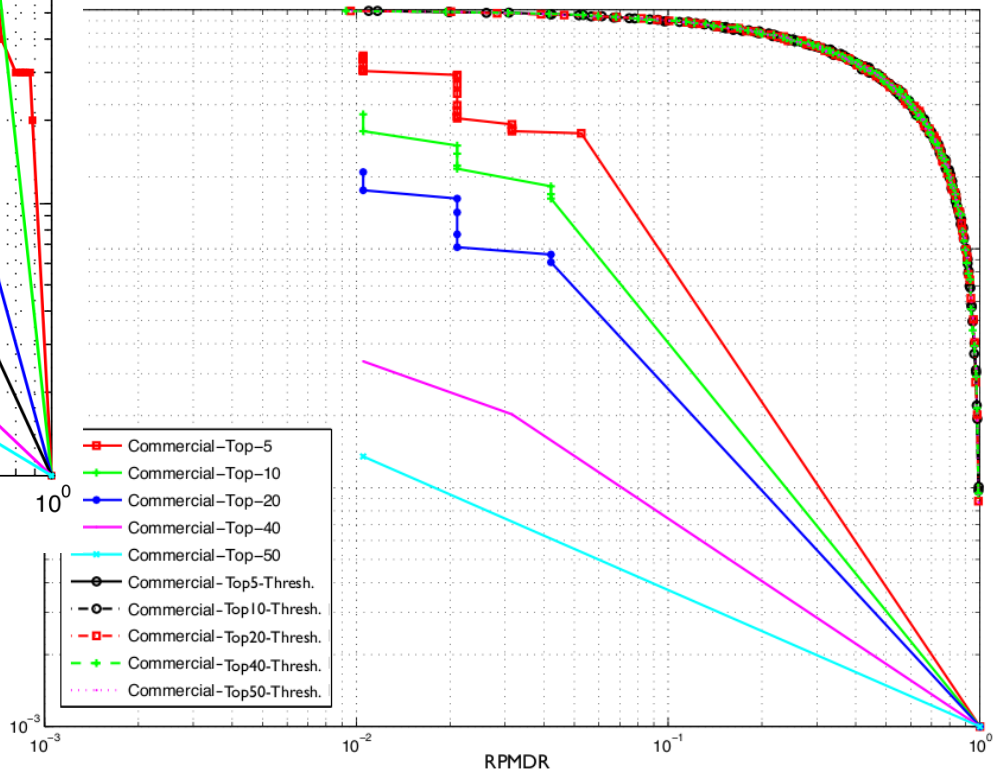
When missing
only 1% of all failures
It has only 10%
false prediction

Rank 1 recognition for
Face recognition algorithm C is 89.4%, 84.5% for face G,
Fingerprint LI 86.5% for and 92.5% for Fingerprint RI

Good failure prediction for all of them, way better than score thresholding

Examined impact of i.i.d. assumptions and sizes of top-N data needed for prediction





What else can Meta-Recognition do?

Decision fusion (fuse only those that are not predicted to fail) or weighted score fusion.

For statistical EVT prediction use:
w-score fusion where:

w-score(x) = CDFWeibull(x)

Use w-score to weight data for fusion, i.e., compute average w-score over different algorithms/modalities.

W. J. Scheirer, A. Rocha, R. J. Micheals, T. E. Boult, “Robust Fusion: Extreme Value Theory for Recognition Score Analysis,” ECCV 2010.

w-score normalization

Require: a collection of scores S, of vector length m, from a single recognition algorithm j;

1. Sort and retain the n largest scores, $s_1, \ldots, s_n \in S$;

2. Fit a Weibull distribution W_S to $s_2, \ldots, s_n$, skipping the hypothesized outlier;

3. While $k < m$ **do**

4. $s'_k = \mathrm{CDF}(s_k, W_S)$

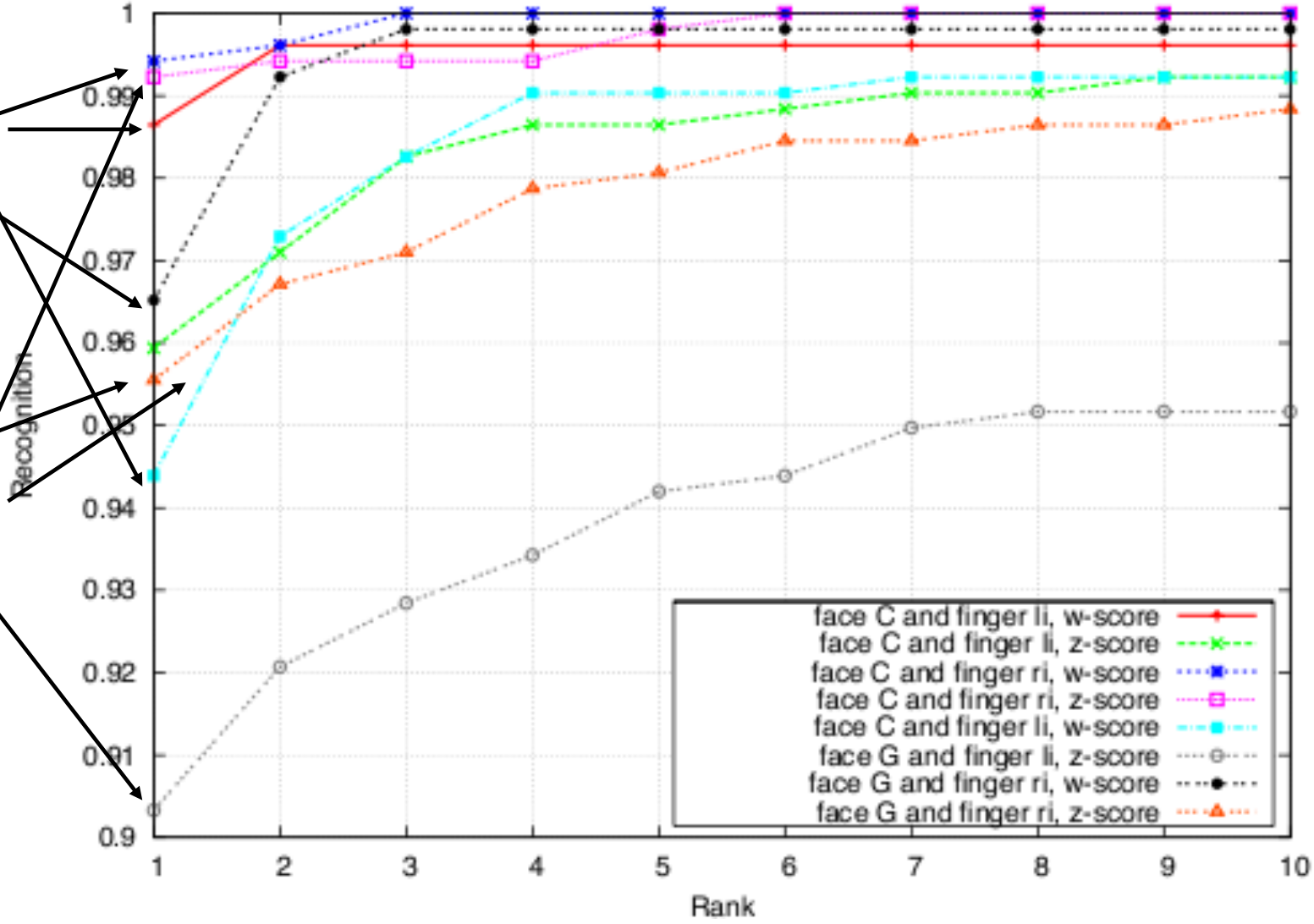
5. $k = k + 1$

6. end while

Fusion Performance

w-score

z-score



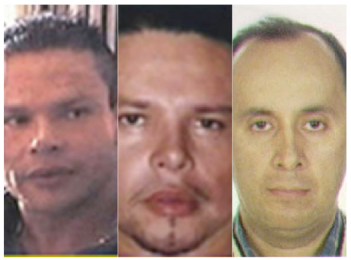
w-score fusion outperformed z-score with sum (or product) fusion on all experiments. In general, the lower the performance the greater the differential gain.

Fusion Problems For:

- Existing theories for fusion algorithms presume consistent data and work to address noise. What happens when user intentionally attempts to thwart the system by changing/destroying their data?
- What is needed is an approach to predict when a particular modality/algorithm is failing and then ignore it.

Failure Prediction and Fusion

Traditional Fusion can be degrade system performance, especially when adversaries try to defeat it.



→ Face Recognition Algorithm → Rejected (not in DB) ✗

Ramirez Abadia

Meta-recognition can predict and select correct modality.

→ Voice Recognition Algorithm → Recognized (as Ramirez Abadia) ✓

Meta-recognition is a useful mathematical theory for fusion that predicts “failing” data

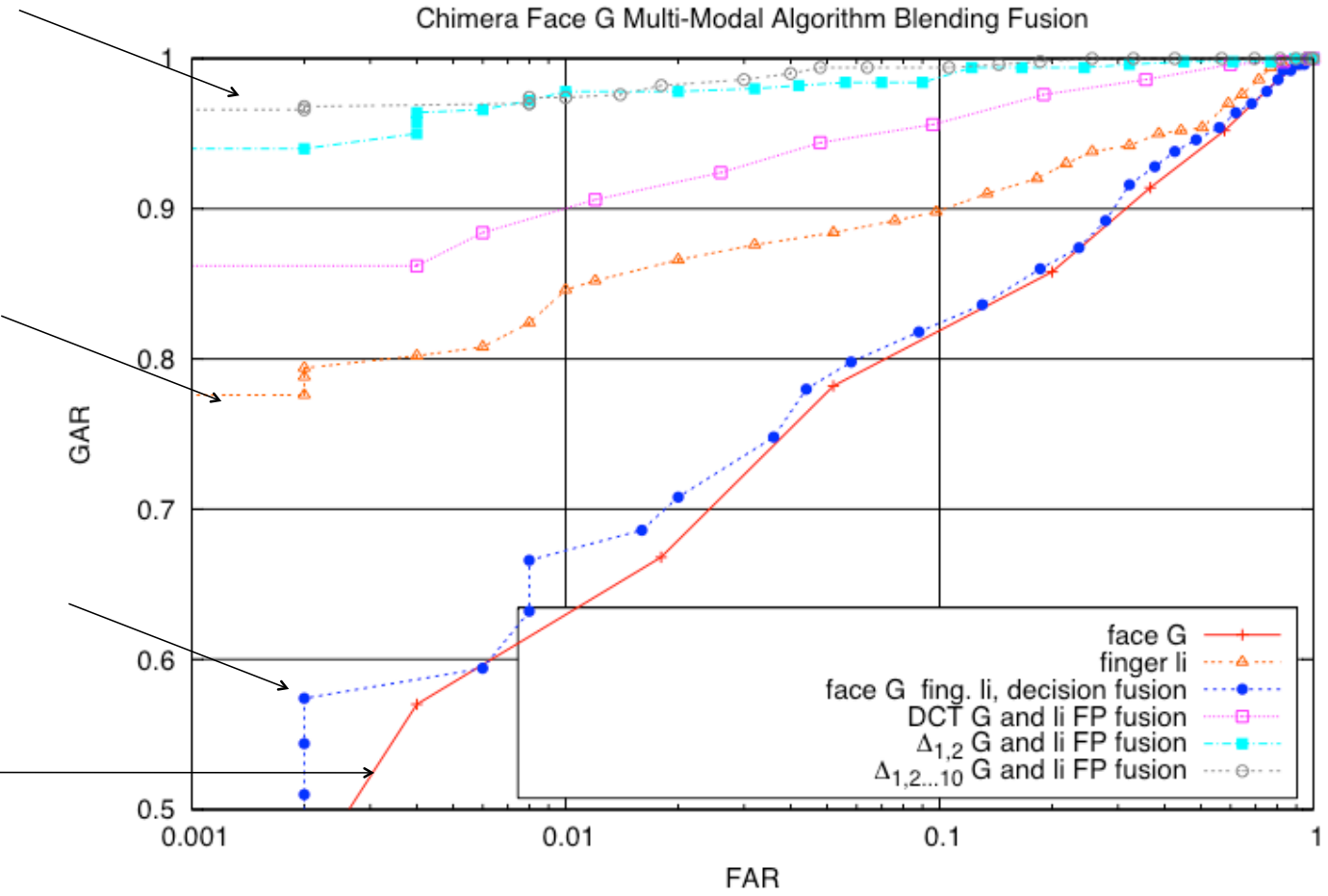
Classic fusion can make things worse!

MR-L fusion for
Face G and Finger Li
Using top-10 matches

Raw Finger Li

Classic "Decision" fusion
for Face G and Finger Li

Raw Face G



BSSR1 has only 600 paired sets of data and was too easy. So we made chimera data, mixing all fingers and faces (6000 samples).

This shows the real power of MR– automatically ignoring bad data!

Failure-prediction W-score fusion vs z-score

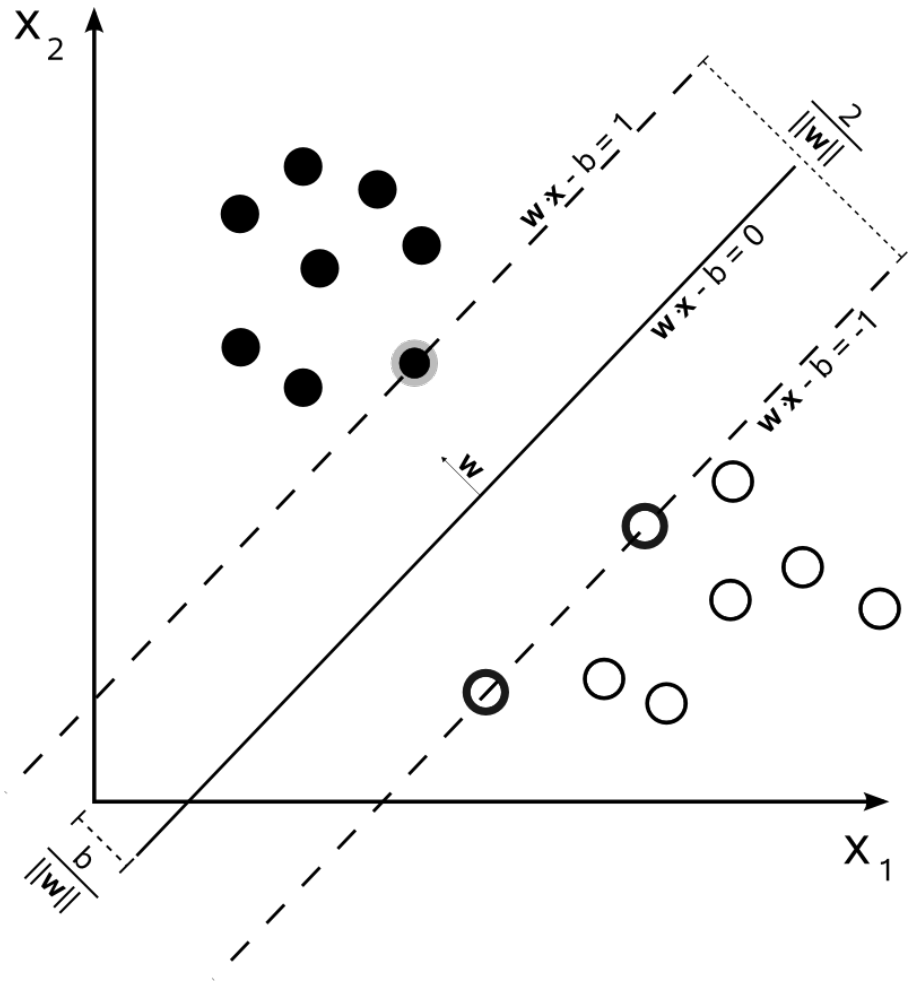
W	Z	
81.6	65.2	face C (impostor), finger LI:
88.1	67.4	face C (impostor), finger RI:
81.6	65.9	face G (impostor), finger LI:
88.1	68.1	face G (impostor), finger RI:
73.3	58.0	face C (impostor), face G (impostor), finger LI:
79.8	60.6	face C (impostor), face G (impostor), finger RI:

3000 samples from NIST BSSR1 data

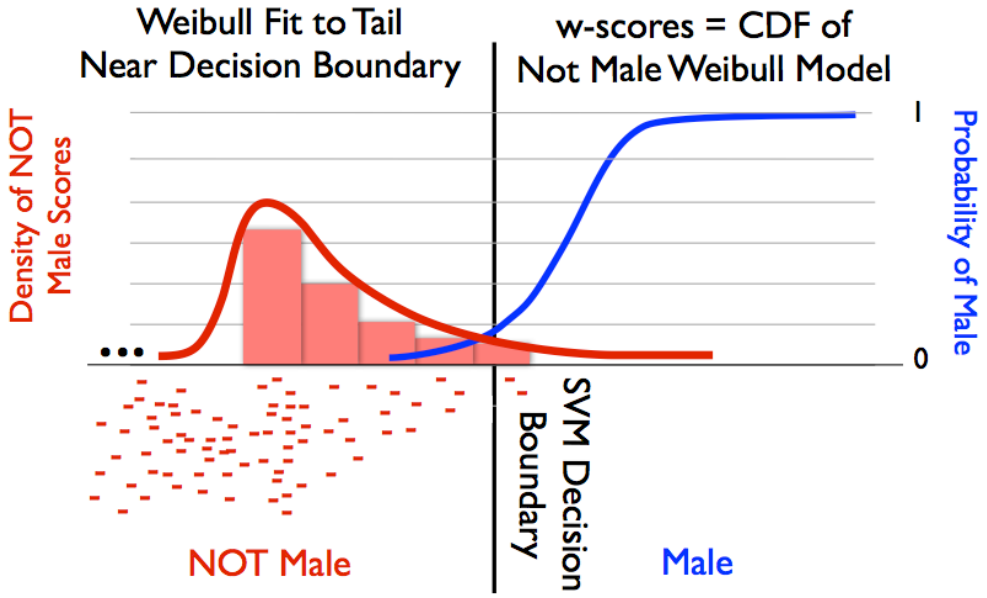
Rank 1 fusion with z-scores is highly impacted by failing modalities; the failure prediction fusion with w-scores is very close to rank 1 of the modality that isn't failing, even with multiple failures.

Support Vectors

- Probability calibration is only well defined close to the decision boundary (Bartlett and Tewari JMLR 2007)
- Boundary is defined by the training samples that are effectively extremes,
  - Calibration models should be based on EVT



Calibration for decision boundaries



1. Get tail of decision scores from the **opposite** class
2. Fit **Weibull** to values in the **tail:**

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

3. Compute normalized scores using **CDF** of the Weibull:

$$F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k}$$

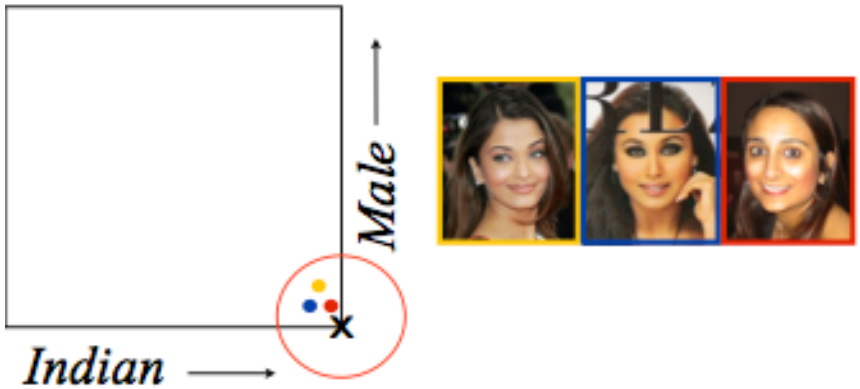
W. J. Scheirer, N. Kumar, P. N. Belhumeur, T. E. Boult, "Multi-Attribute Spaces: Calibration for Attribute Fusion and Similarity Search," CVPR 2012

Fusion after normalization

1. maximize over I	$s^q = \| A_j(I) \|_1$	The goal is to find images I that maximize the L_1 norm for each attribute j in the query set J
2. subject to	$A_j(I) = F(T(s_j(I)); W_j)$	
3. for $\forall j \in J$ satisfying	$0 \leq \alpha_j \leq A_j(I) \leq \beta_j \leq 1$	

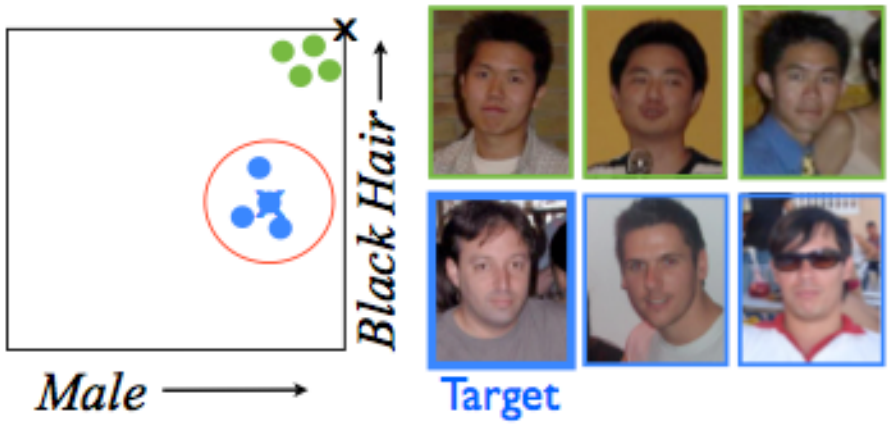
Multi-Attribute Search

"Indian Females"

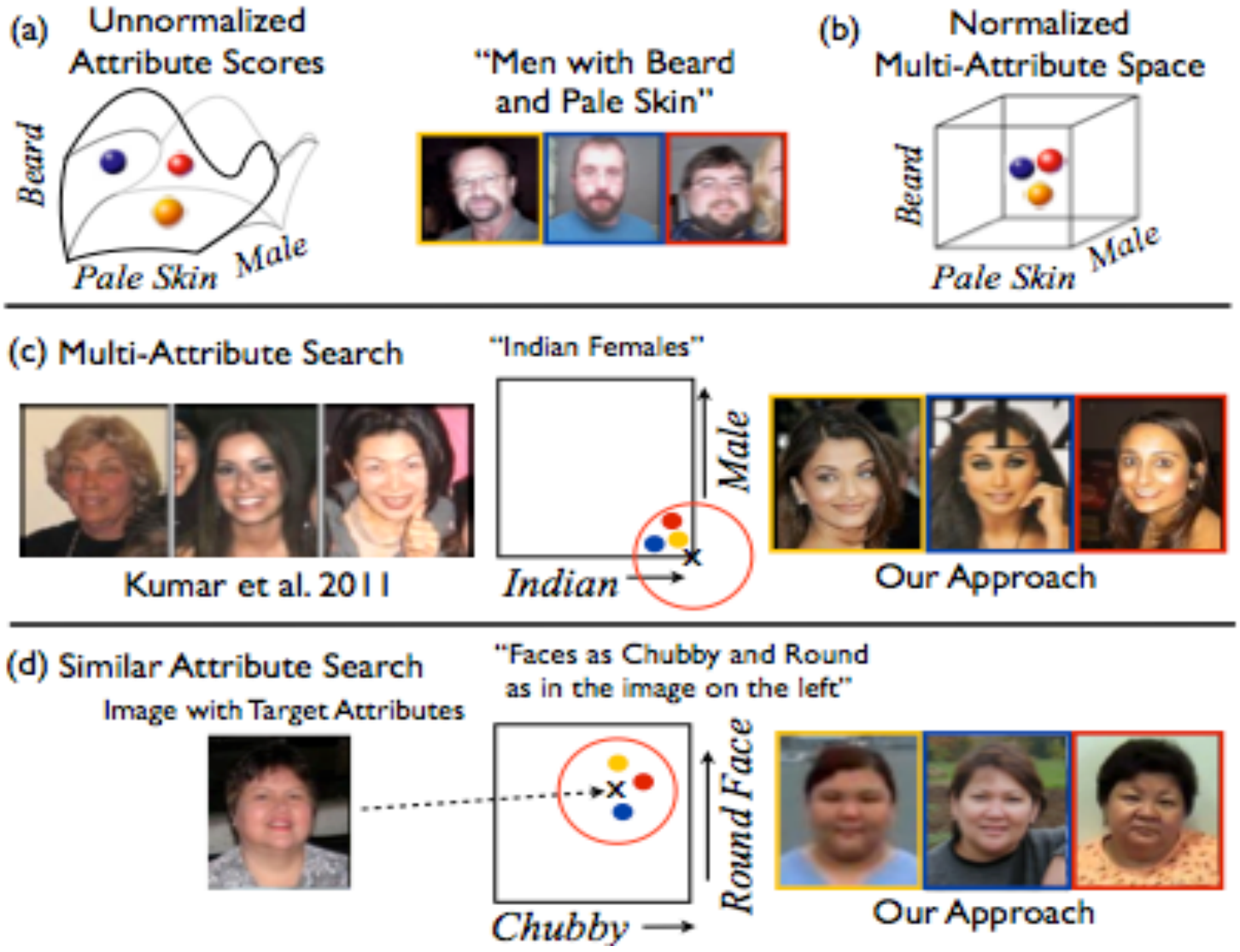


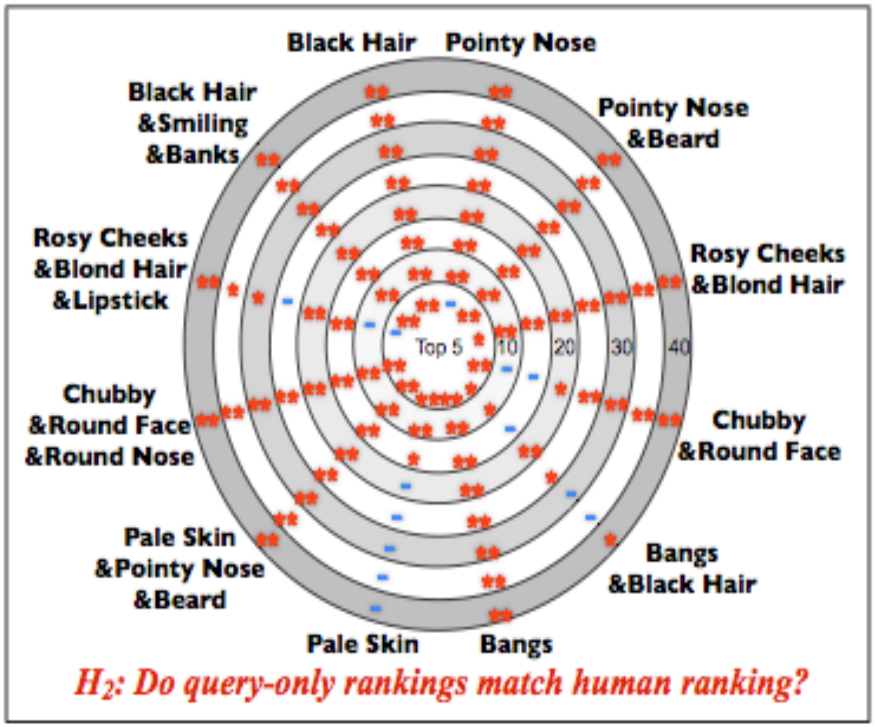
Target Attribute Similarity Search

"Male and Black Hair Like Target"



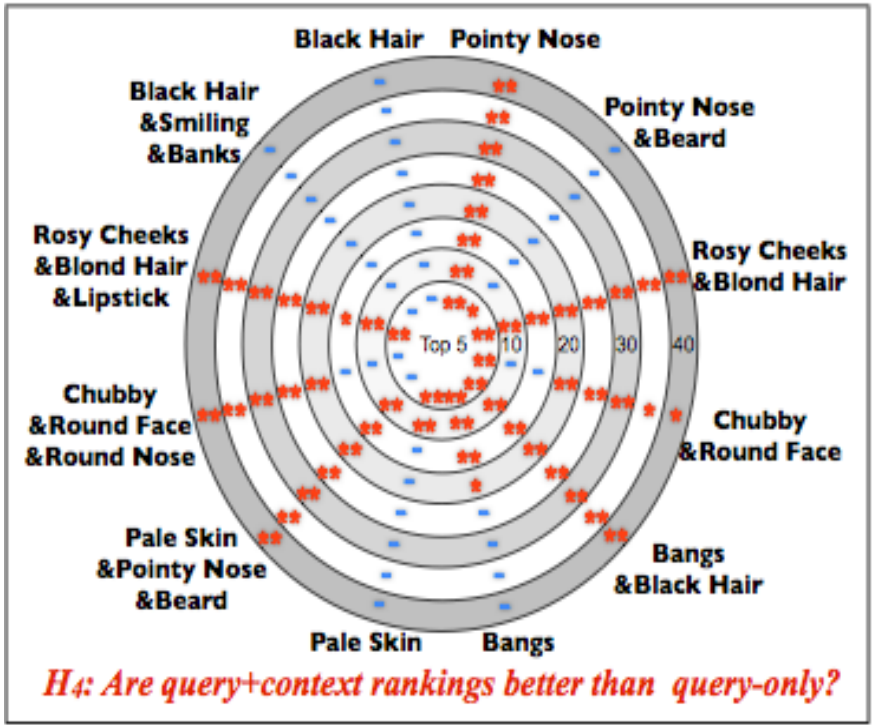
Utility of the calibration model





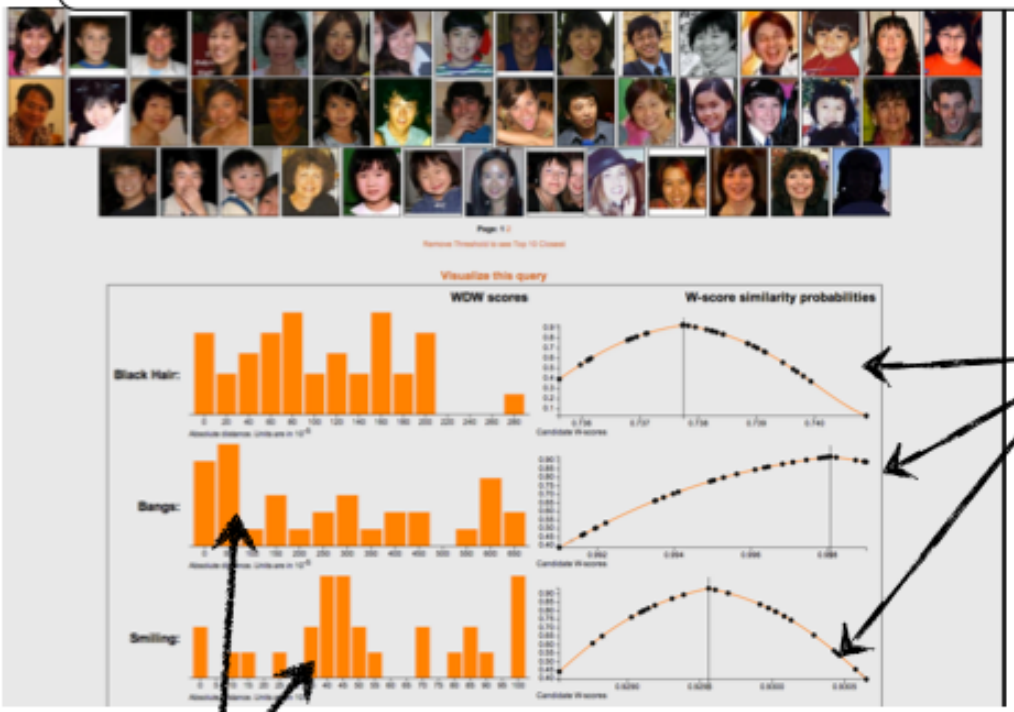
H_2: Do query-only rankings match human ranking?

H_3: Do query+context rankings match human ranking?



H_4: Are query+context rankings better than query-only?

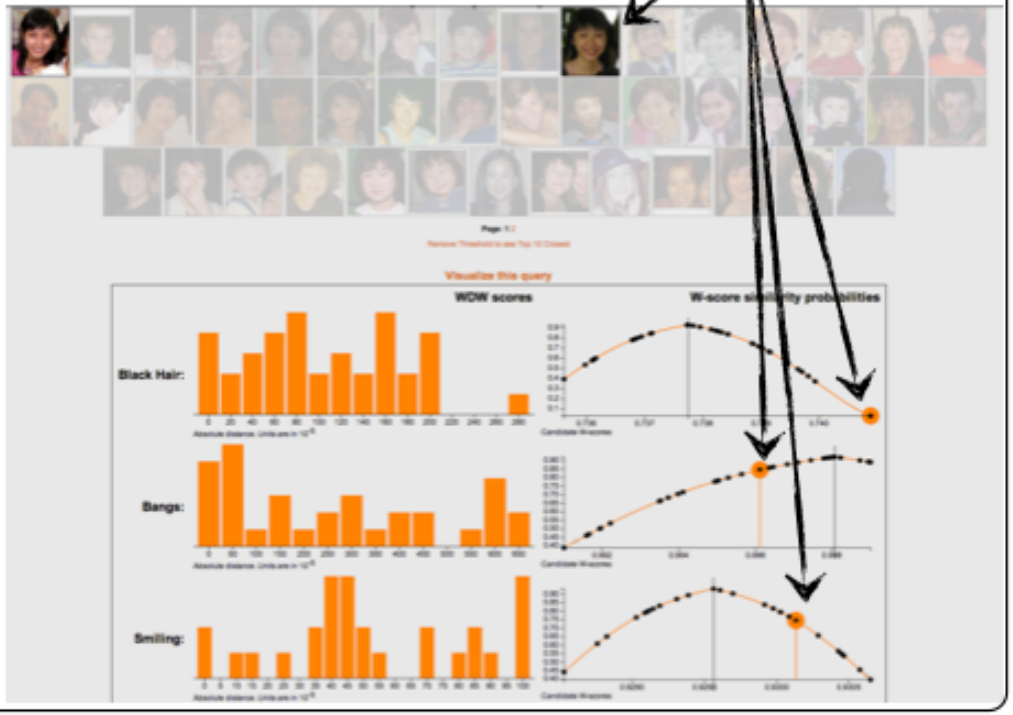
Exploring Similarity Search Results



Per-attribute Similarity (WDW) Score distribution for results of query showing all non-zero scores from top matches

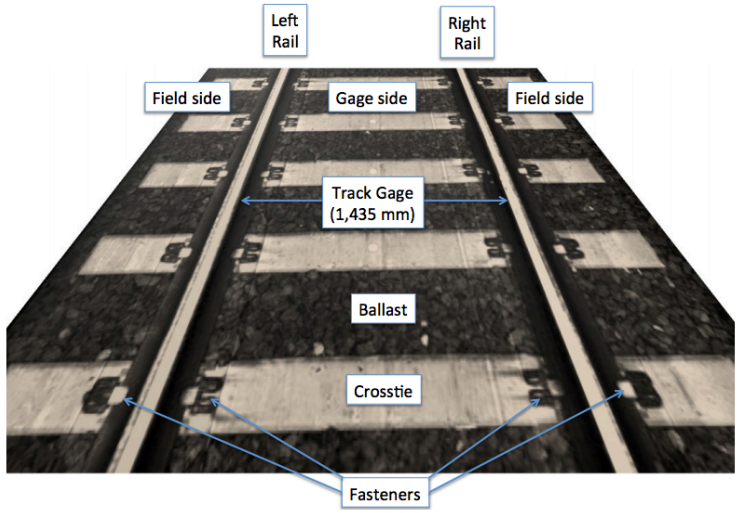
Highlighted query image's similarity score to target (first) image. No highlight means not in top scores

Histograms of absolute difference of w-scores



Note clustering in scores and variations in shape of distributions

Sequential Score Adaptation with Extreme Value Theory



Application: Visual Railway Track Inspection

Same process as the w-score, but swap out the Weibull distribution for the Generalized Pareto Distribution:

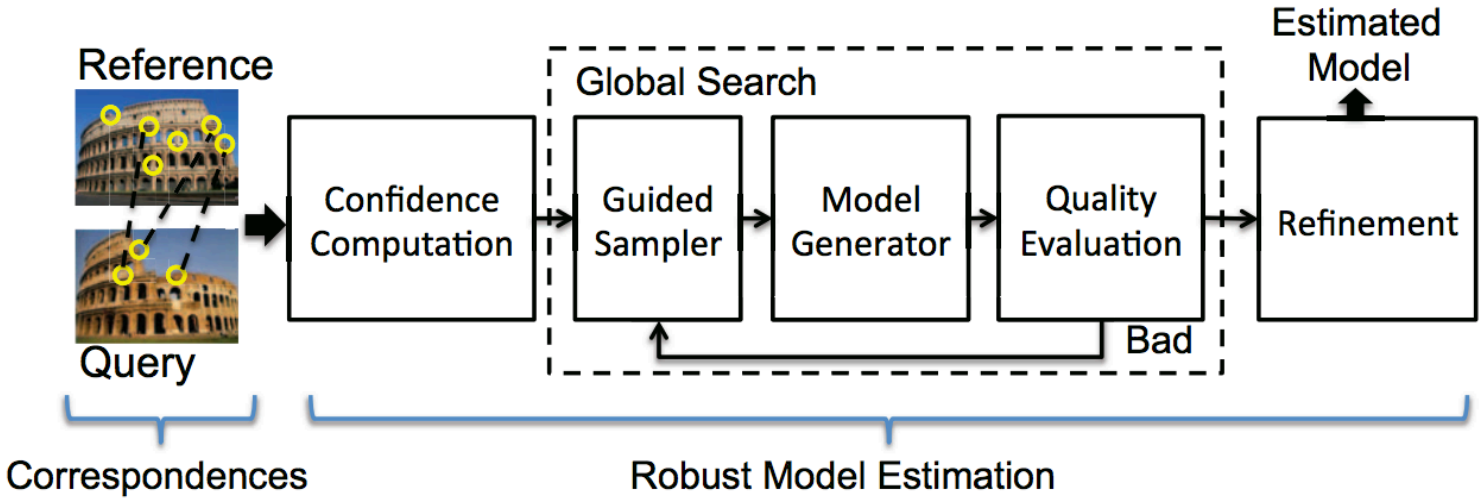
$$G(y;\sigma,\xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \quad y > 0.$$

where $\sigma > 0$, $\xi \in \mathbb{R}$, and $x_+ = \max(x, 0)$.

X. Gibert-Serra, V. M. Patel, and R. Chellappa, "Sequential score adaptation with extreme value theory for robust railway track inspection," Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving (CVRSUAD), Santiago, Chile, 2015.

Sampling and feature correspondence

Fragoso and Turk CVPR 2013



Guided Sampling Methodology with M-R Prediction

V. Fragoso and M. Turk, “SWIGS: A Swift Guided Sampling Method,” CVPR 2013

M-R Rayleigh

Same meta-recognition algorithm, but constrain the Weibull distribution to be Rayleigh distribution:

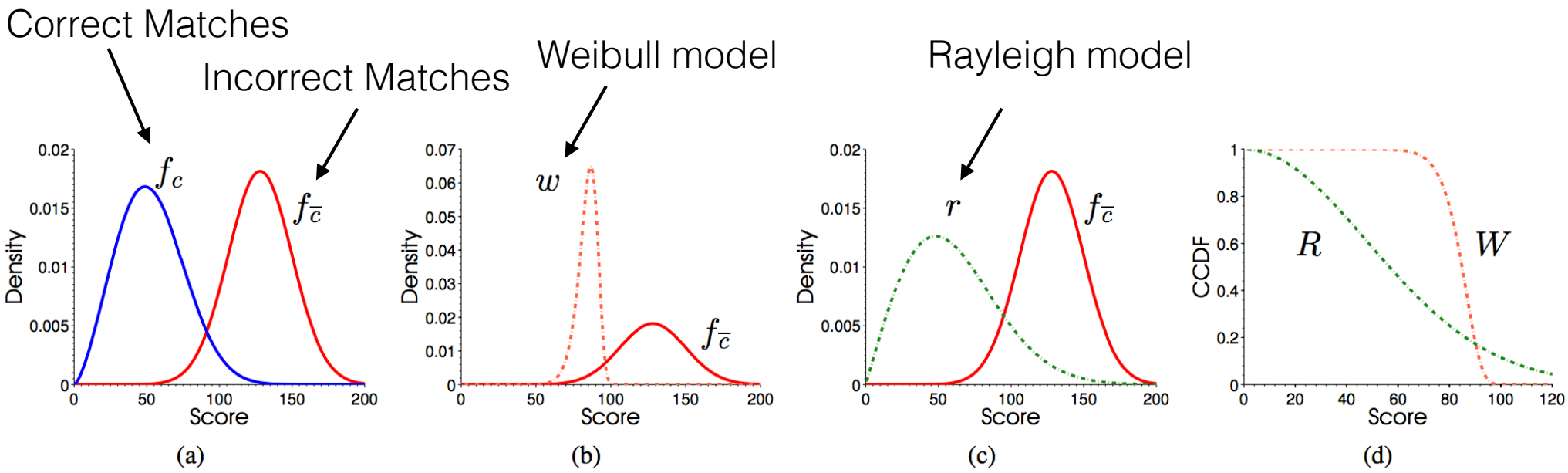
$$R(s;\sigma) = e^{-\frac{s^2}{2\sigma^2}},$$

Advantage: one parameter to fit

Estimate σ from the closest scores $s_{2:k}$ using the maximum-likelihood formula:

$$\hat{\sigma} = \sqrt{\frac{1}{2\,(k-1)} \sum_{j=2}^{k} s_j^2}$$

Densities involved in a keypoint matching process per query frame



M-R Rayleigh vs. M-R Weibull

Feature correspondences
Top: SIFT, Bottom: SURF