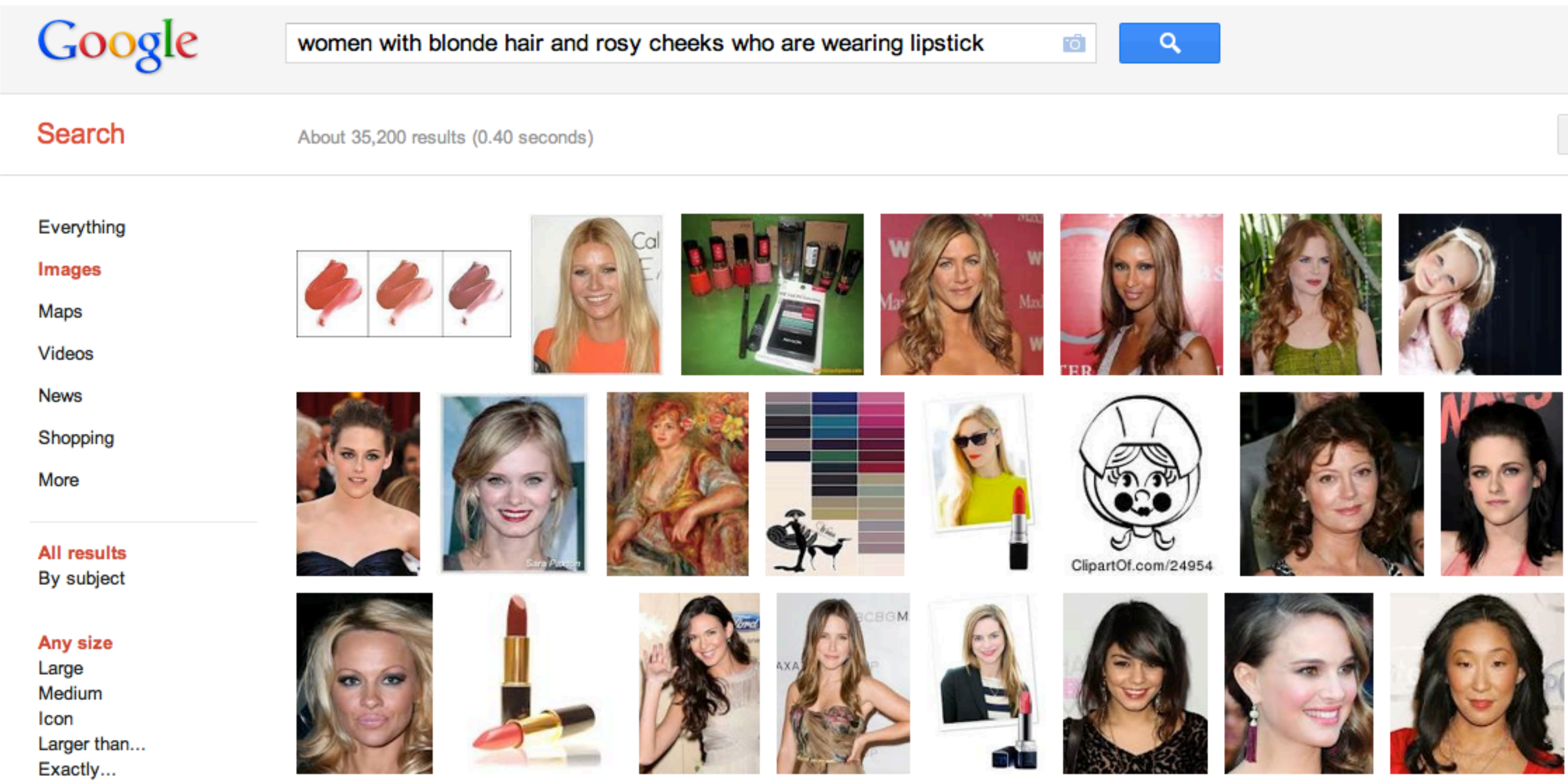
Meta-Recognition: Score Analysis and Calibration for Recognition Problems

Walter J. Scheirer
Vision and Security Technology Lab
University of Colorado at Colorado Springs

How can we find images of women with blonde hair and rosy cheeks who are wearing lipstick?

What do we get with the most popular image retrieval tool?



Visual Attributes

- Ferrari and Zisserman NIPS 2007[1]
  - Describe objects by their *attributes*

Has Horn
Has Leg
Has Head
Has Wool

textual description

Mountain Goat (cc) by-nc-nd Cliff Hall

- Kumar et al. T-PAMI 2011[2]
  - Describe faces by their *attributes*

Has Hat
Has Beard
Has African Ethnicity
Has Round Nose

textual description

Ghostface Killah (cc) by-nc-nd Enrico Fuente

1. V. Ferrari and A. Zisserman, "Learning Visual Attrivutes," NIPS 2007
2. N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Describable Visual Attributes for Face Verification and Image Search," IEEE T-PAMI, 2011

Visual Facial Attributes

- Kumar et al. 2011
  - Low-level simple features + machine learning
    - Feature extractors are composed of pixels from face region, pixel feature type, normalization and aggregation
    - From an aligned image I, extract low level features:

$$\mathcal{F}(I) = \{\mathbf{f}_1(I), \ldots, \mathbf{f}_k(I)\}$$

    - In total, we trained **73** different SVM attributes classifiers
    - Crowdsourced ground truth labeling; 500-2000 +/- examples from the Columbia Face Database

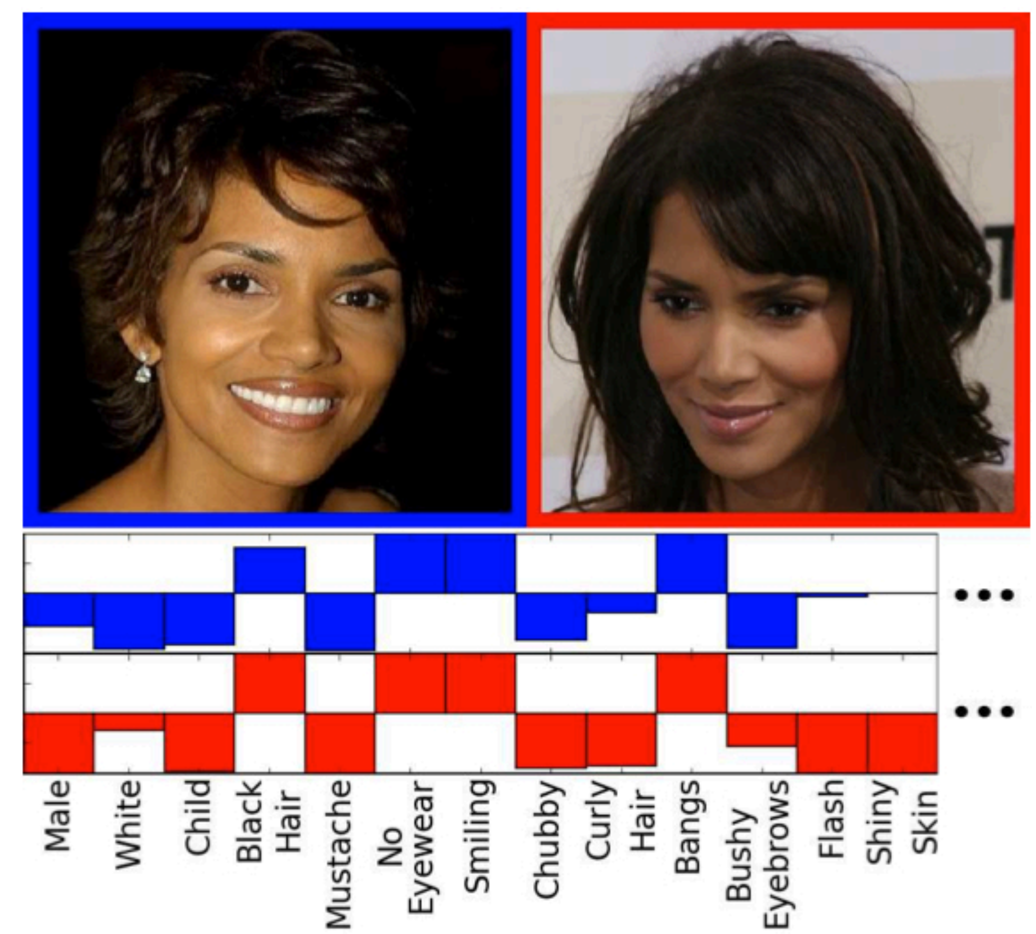


Image adapted from Fig 1. in N. Kumar et al. "Describable Visual Attributes for Face Verification and Image Search," *T-PAMI*, 2011

We can use combinations of attributes for search

Search Query: **White Babies Wearing Hats**

Results Produced by the approach of Kumar et al. in T-PAMI 2011

But what's the problem here?

Let's try to build a **multi-attribute space**[1] through the calibration of SVM decision scores

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

How does it work?

The calibration of the decision scores from a binary SVM can be accomplished through the use of **Meta-Recognition**.

Our robust normalization converts the decision scores to **w-scores**, which are estimated probabilities of an attribute NOT being drawn from the class opposite to it.

A **multi-attribute space** is a product space formed from well normalized attribute functions.

What is recognition in computer vision?

- Compare an object to a known set of classes, producing a similarity measure to each

Teapot

Red teapot (cc) by fraise

Frog

Frog on corn leaf (cc) by Joi Ito

Quiet brown frog (cc) by Olivier Ffrench

What is this?

Girl

Lovely little girl:) (cc) by BirdCantFly

Data Fusion

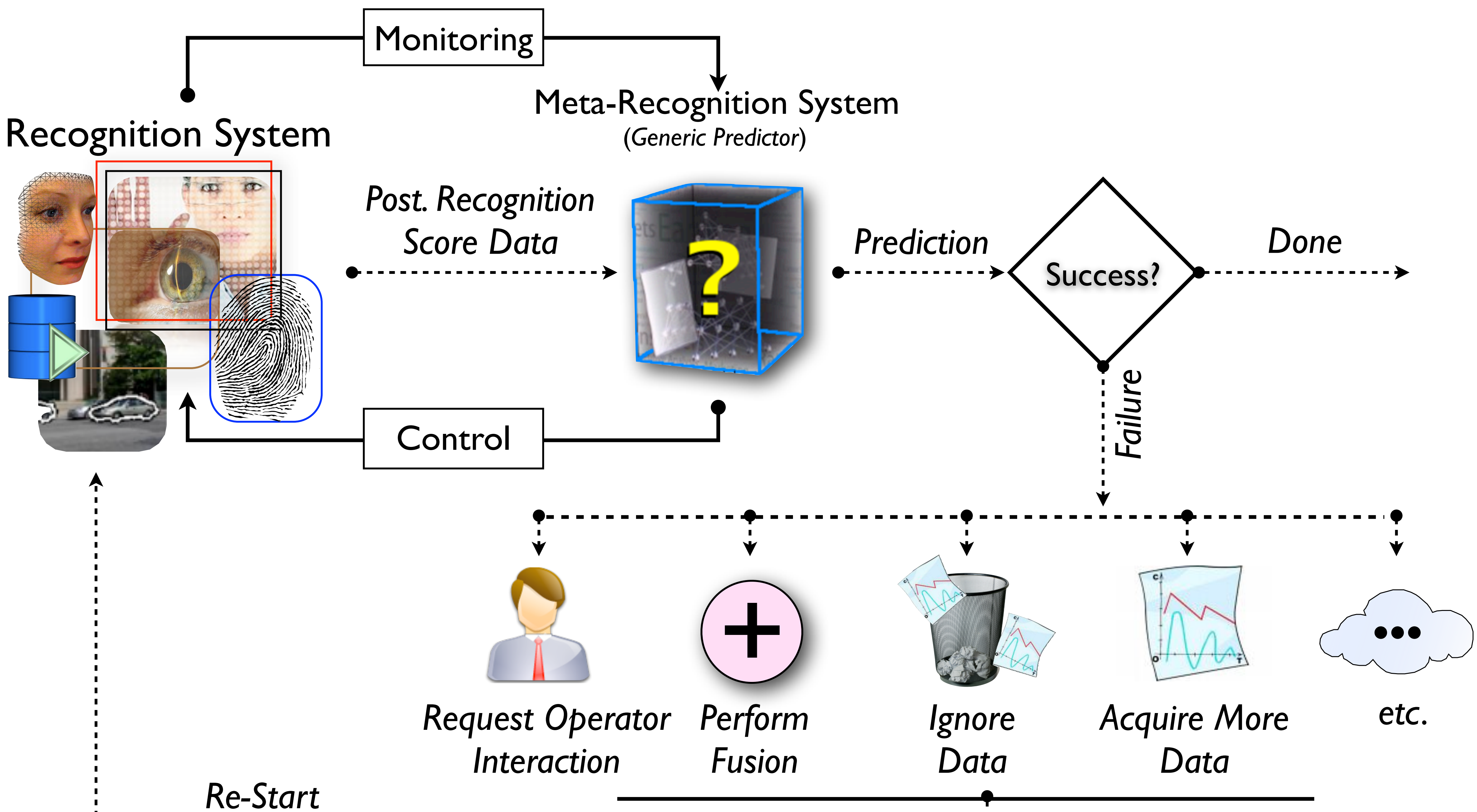
- A single algorithm is not a complete solution for a recognition task
- Combine information across algorithms, classifiers, or sensors[1]
  - Decision fusion
  - Score level normalization & fusion

Do this is a **robust** manner...

1. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, 2006

Meta-Recognition

Goal: Predict if a recognition result is a success or failure



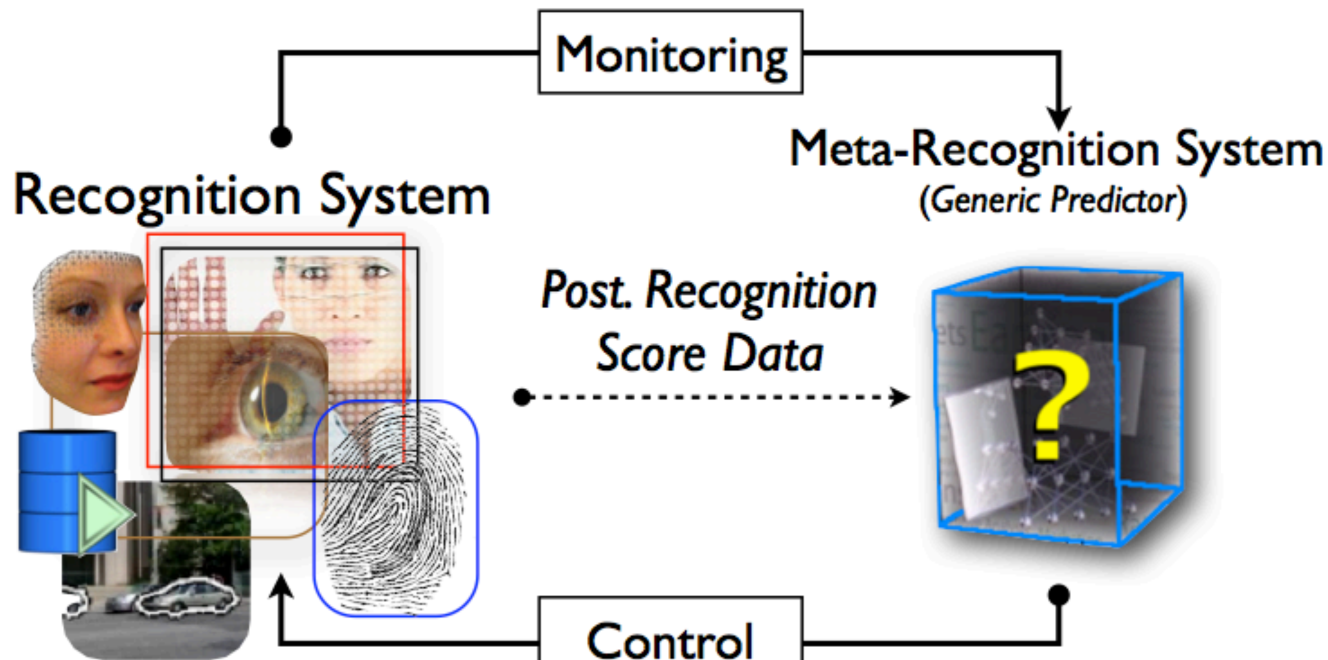
W. Scheirer et al., "Meta-Recognition: the Theory and Practice of Recognition Score Analysis," IEEE T-PAMI, August 2011

From Meta-Cognition to Recognition

- Inspiration: *Meta-Cognition* Study
  - "knowing about knowing[1]"
  - Example: If a student has more trouble learning history than math, she "knows" something about her learning ability and can take corrective action

1. J. Flavell and H. Wellman, "Metamemory," in *Perspectives on the Development of Memory and Cognition*, 1988, pp. 3-33

Meta-Recognition Defined



Let X be a recognition system. Y is 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 on the observations.

Can’t we do this with say... image quality?

8 47

191 Gallery

Apparent quality is not always tied to rank.

- Quality is good as an “overall” predictor
  - Over a large series of data and time

- Quality does not work as a “per instance” predictor
  - One image analyzed at a time...

Challenges for Image Quality Assessment

- Interesting recent studies from the National Institute of Standards and Technology
  - Iris[1]: three different quality assessment algorithms lacked correlation
  - Face[2]: out of focus imagery was shown to produce better match scores

"Quality is not in the eye of the beholder; it is in the recognition performance figures!" - Ross Beveridge

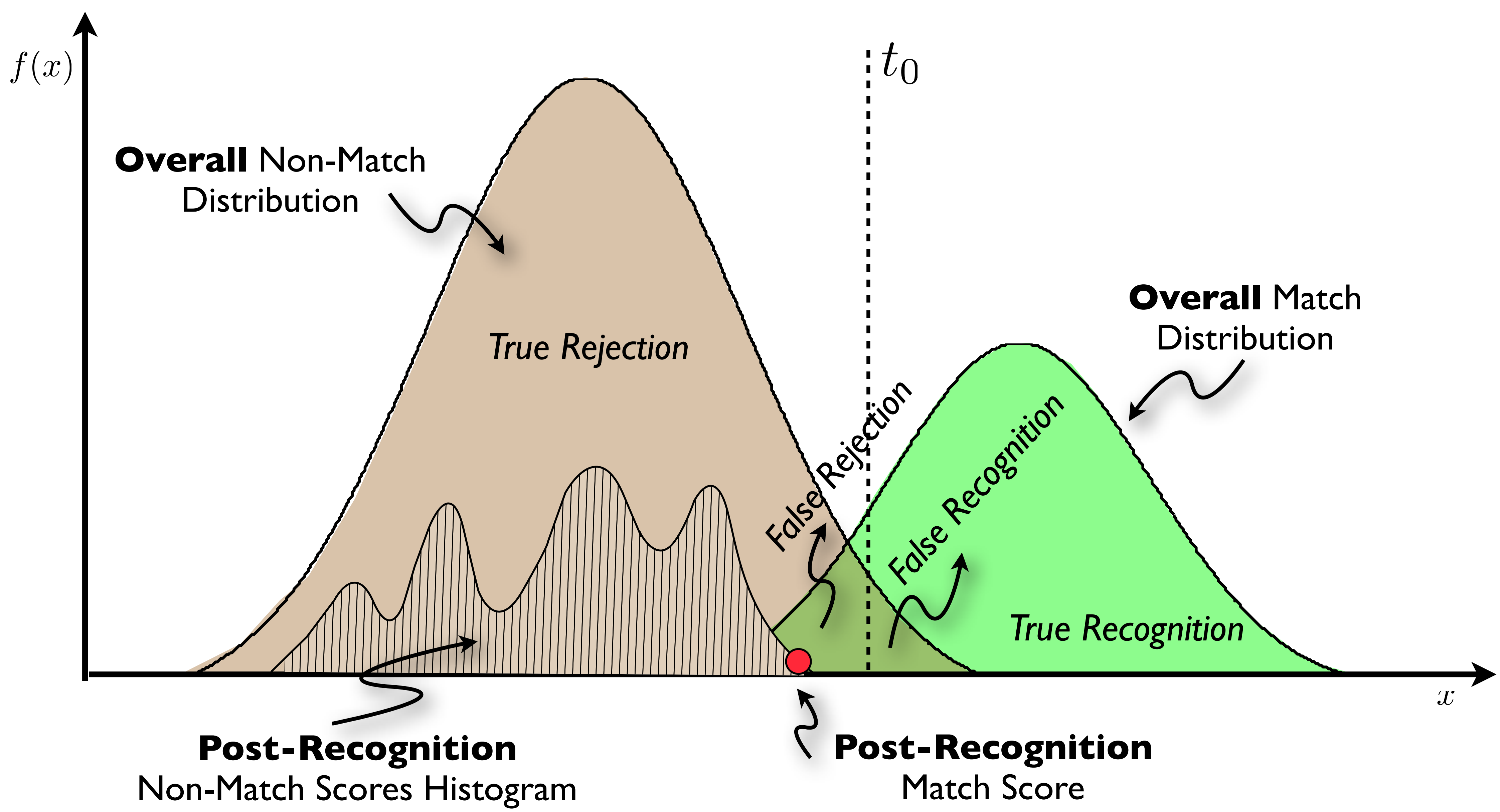
1. P. Flynn, "ICE Mining: Quality and Demographic Investigations of ICE 2006 Performance Results," MBGC Kick-off workshop, 2008
2. R. Beveridge, "Face Recognition Vendor Test 2006 Experiment 4 Covariate Study," MBGC Kick-off workshop, 2008

What about cohorts?

- A likely related phenomenon to Meta-Recognition
- Post-*verification* score analysis
- Model a distribution of scores from a pre-defined "cohort gallery" and then normalize data[1]
  - This estimate valid "score neighbors"
  - A claimed object should be followed by its cohorts with a high degree of probability
- Intuitive, but lacks a theoretical basis

1. S. Tulyakov et al., "Comparison of Combination Methods Utilizing t-normalization and Second Best Score Models," IEEE Workshop on Biometrics, 2008.

Recognition Systems



Formal definition of recognition

Find[1] the class label c^*, where p_k is an underlying probability rule and p_0 is the input distribution satisfying:

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

subject to $\Pr(p_0 = p_{c^*}) \geq 1 - \delta$, for a given confidence threshold δ. We can also conclude a lack of such class.

Probe: input image p_0 submitted to the system with corresponding class label c^*.

Gallery: all the classes c^* known by the recognition system.

1. G. Shakhnarovich, et al. "Face Recognition from Long-term Observations," ECCV, 2002.

Rank-1 Prediction as a Hypothesis Test

- Formalization of Meta-Recognition
  - Determine if the top K scores contain an outlier with respect to the current probe's match distribution
- Let $\mathcal{F}(p)$ be the non-match distribution, and $m(p)$ be the match score for that probe.
- Let $S(K) = s_1 \ldots s_k$ be the top K sorted scores

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

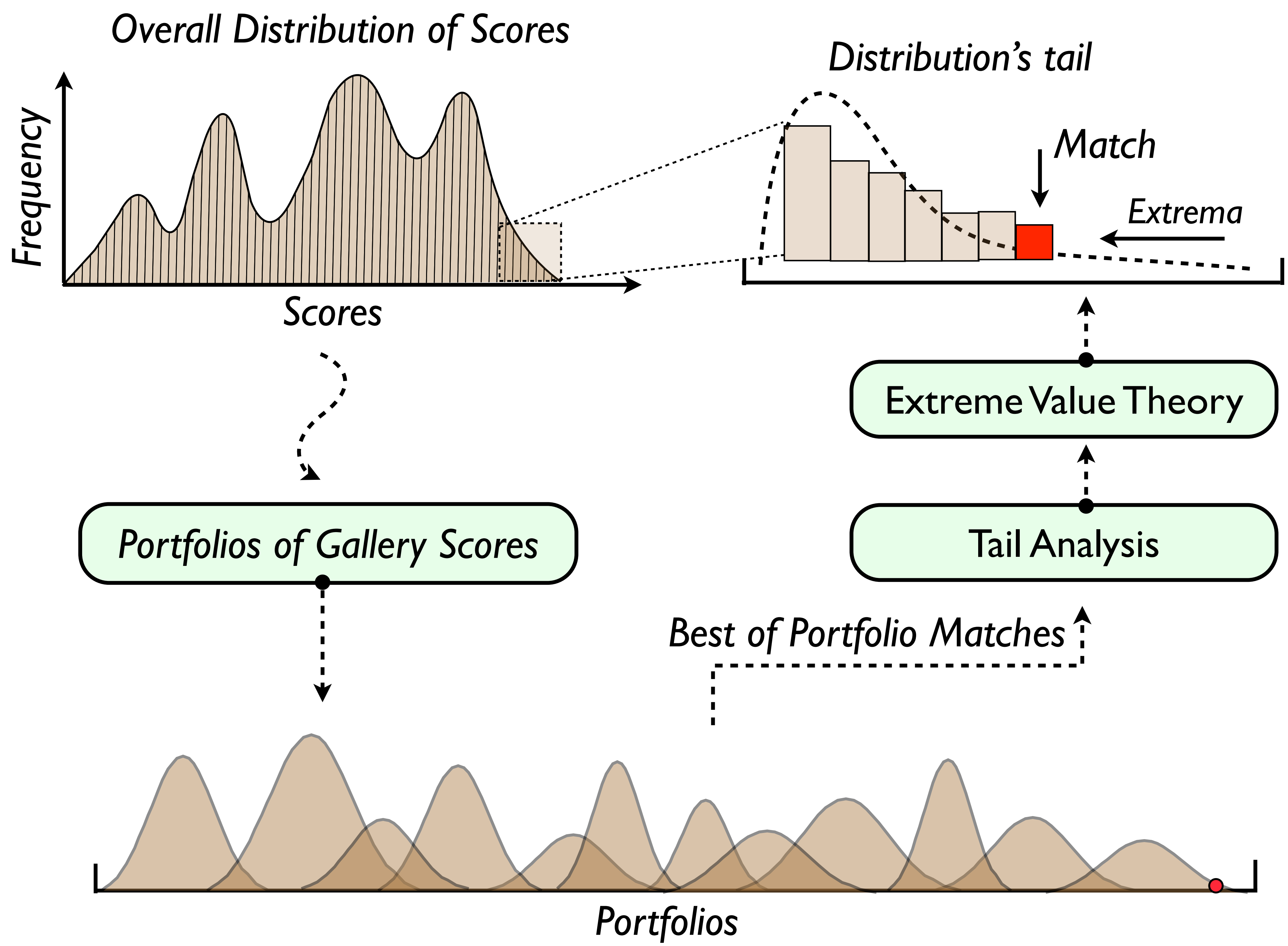
If we can reject H_0, then we predict success.

The Key Insight

We don't have enough data to model the match distribution, but we have n samples of the non-match distribution - good enough for non-match modeling and outlier detection.

If the best score is a match, then it should be an outlier with respect to the non-match model.

A Portfolio Model of Recognition



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

Rank-1 Statistical Meta-Recognition

Require: a collection of similarity scores S

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

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

3. **if** $Inv(W_S(s_1)) > \delta$ **do**

4. s_1 is an outlier and we reject the failure prediction (null) hypothesis H_0

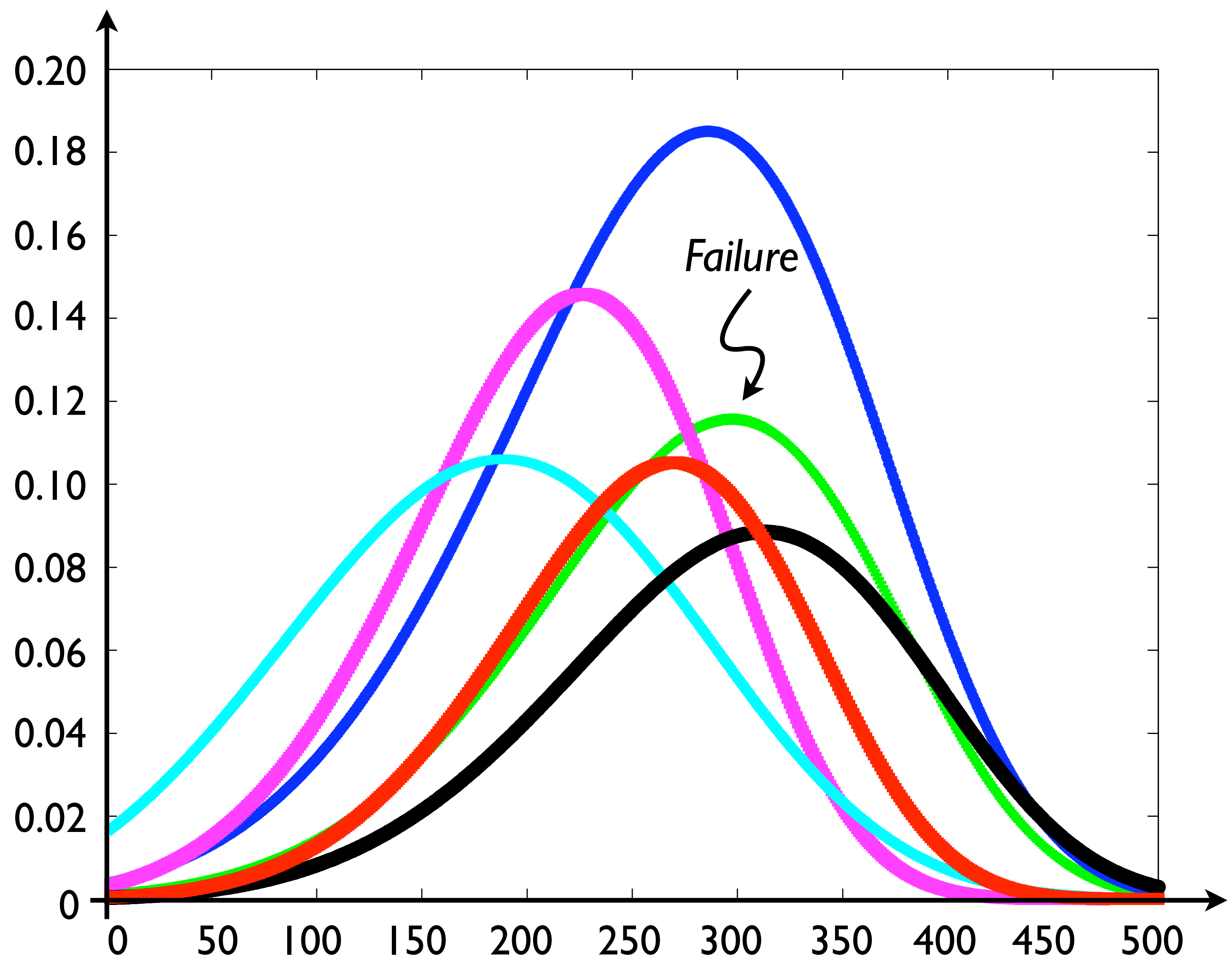
6. **end if**

δ is the hypothesis test "significance" level threshold

Good performance is often achieved using $\delta = 1 - 10^{-8}$

Can't we just look at the mean or shape of the distribution?

Per-instance success and failure distributions are not distinguishable by shape or position



The outlier test is necessary

Meta-Recognition Error Trade-off Curves

	Conventional Explanation	**Prediction**	**Ground Truth**
Case 1	False Accept	Success	O
Case 2	False Reject	Failure	O
Case 3	True Accept	Success	P
Case 4	True Reject	Failure	P

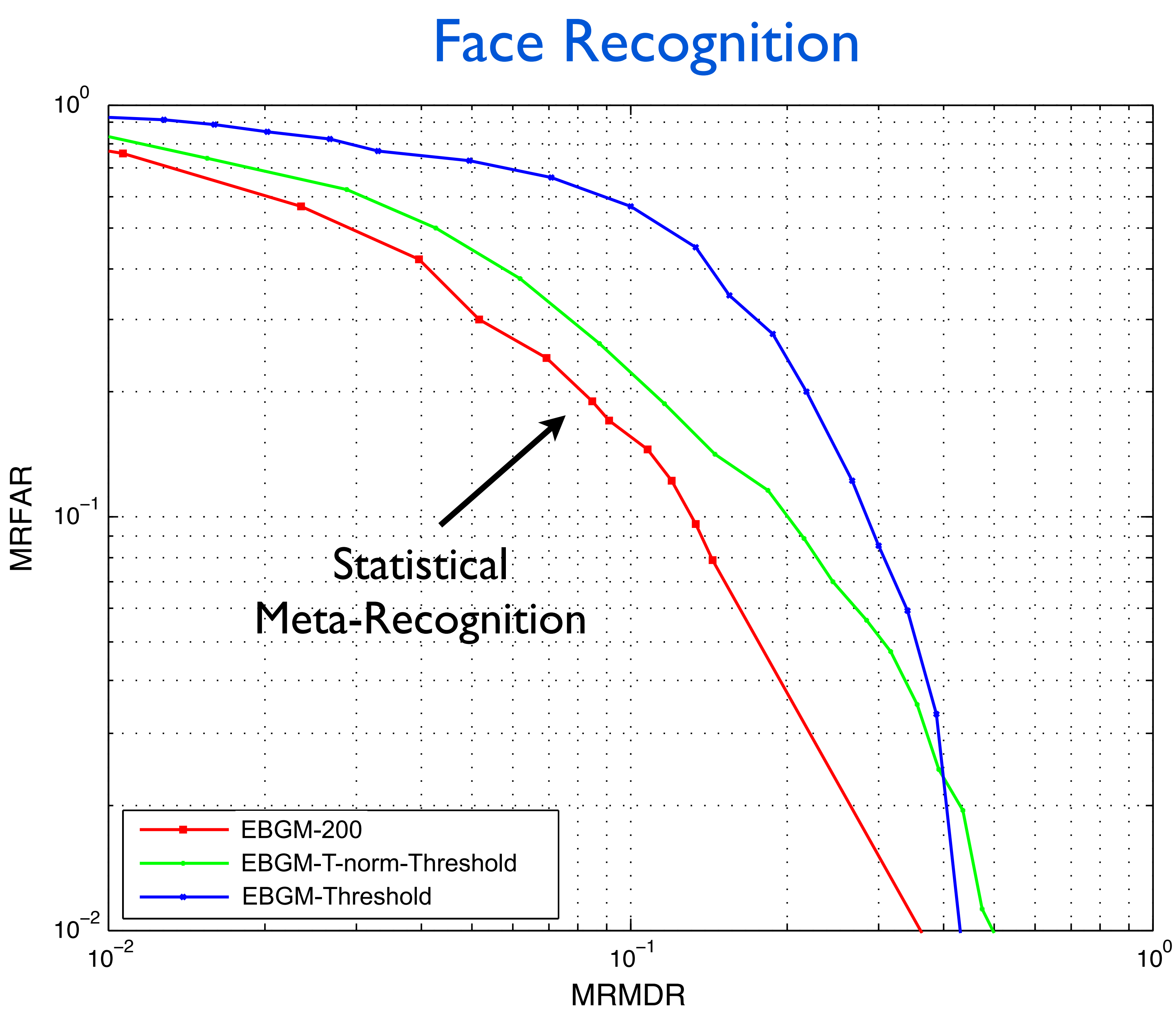
Meta-Recognition False Alarm Rate

$$\text{MRFAR} = \frac{|\text{Case 1}|}{|\text{Case 1}| + |\text{Case 4}|}$$

Meta-Recognition Miss Detection Rate

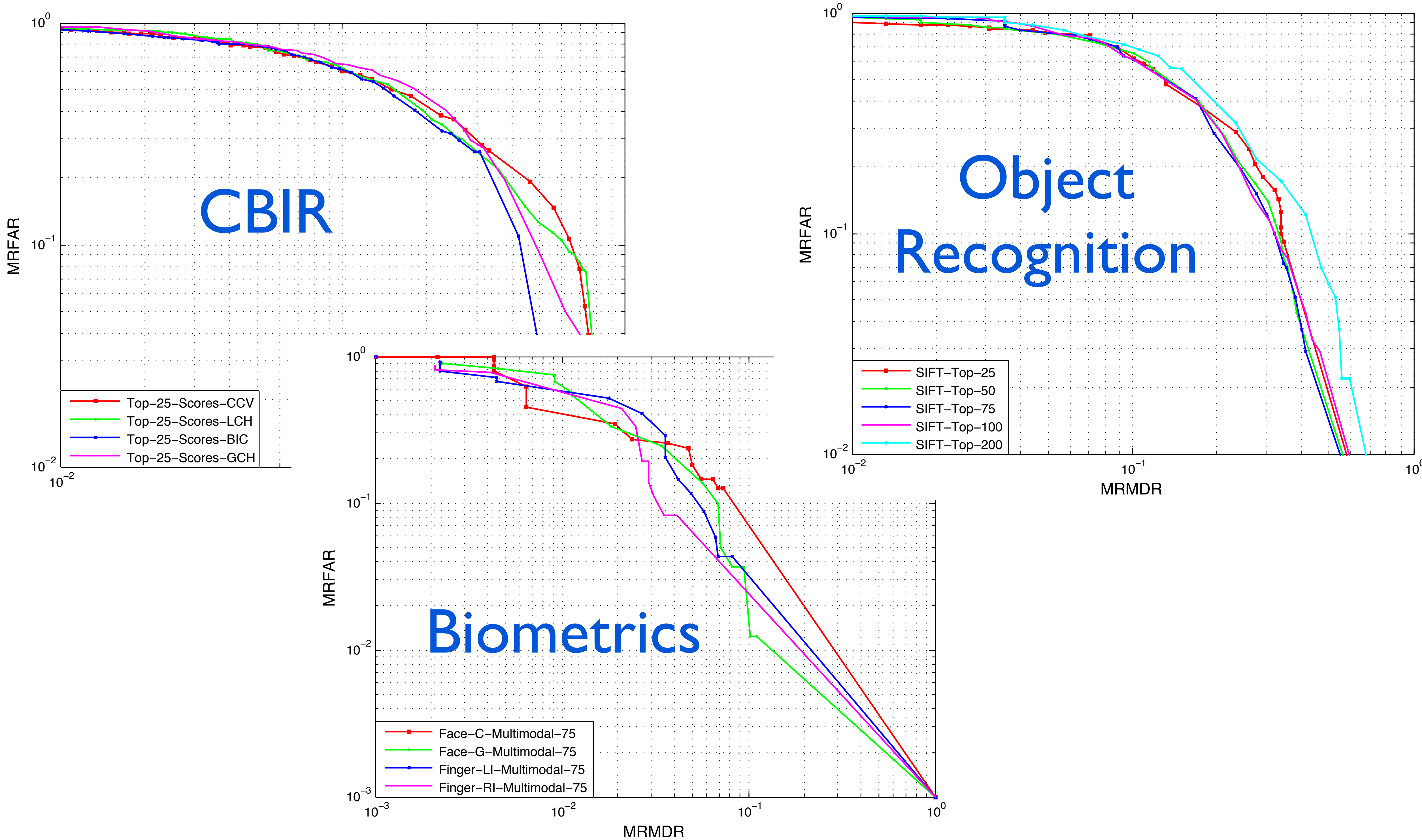
$$\text{MRMDR} = \frac{|\text{Case 2}|}{|\text{Case 2}| + |\text{Case 3}|}$$

Comparison with Basic Thresholding over Original and T-norm Scores



Points approaching the lower left corner minimize both errors

And meta-recognition works across all algorithms tested...

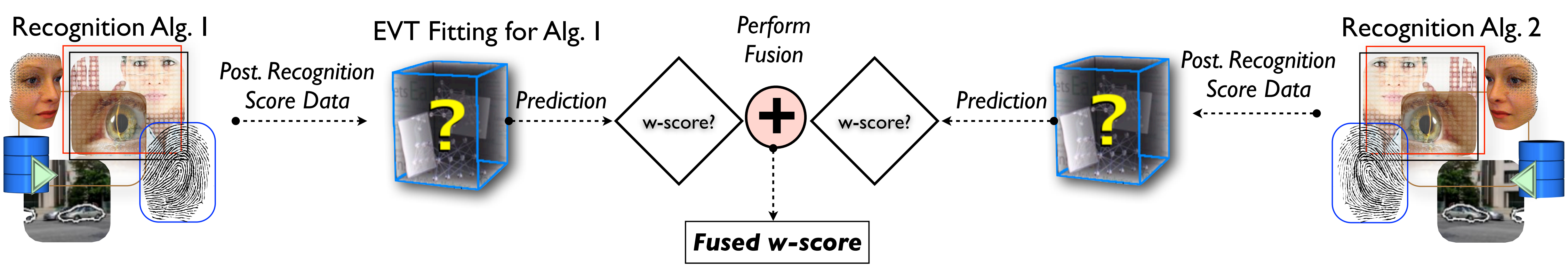


We can do score level fusion too...

Use the CDF of the Weibull model for score normalization:

$$\mathrm{CDF}(x) = 1 - e^{-(x/\lambda)^k}$$

We call this a *w-score*[1]



1. W. Scheirer et al., "Robust Fusion: Extreme Value Theory for Recognition Score Normalization" 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, $s_1, \dots, s_n \in S$;

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

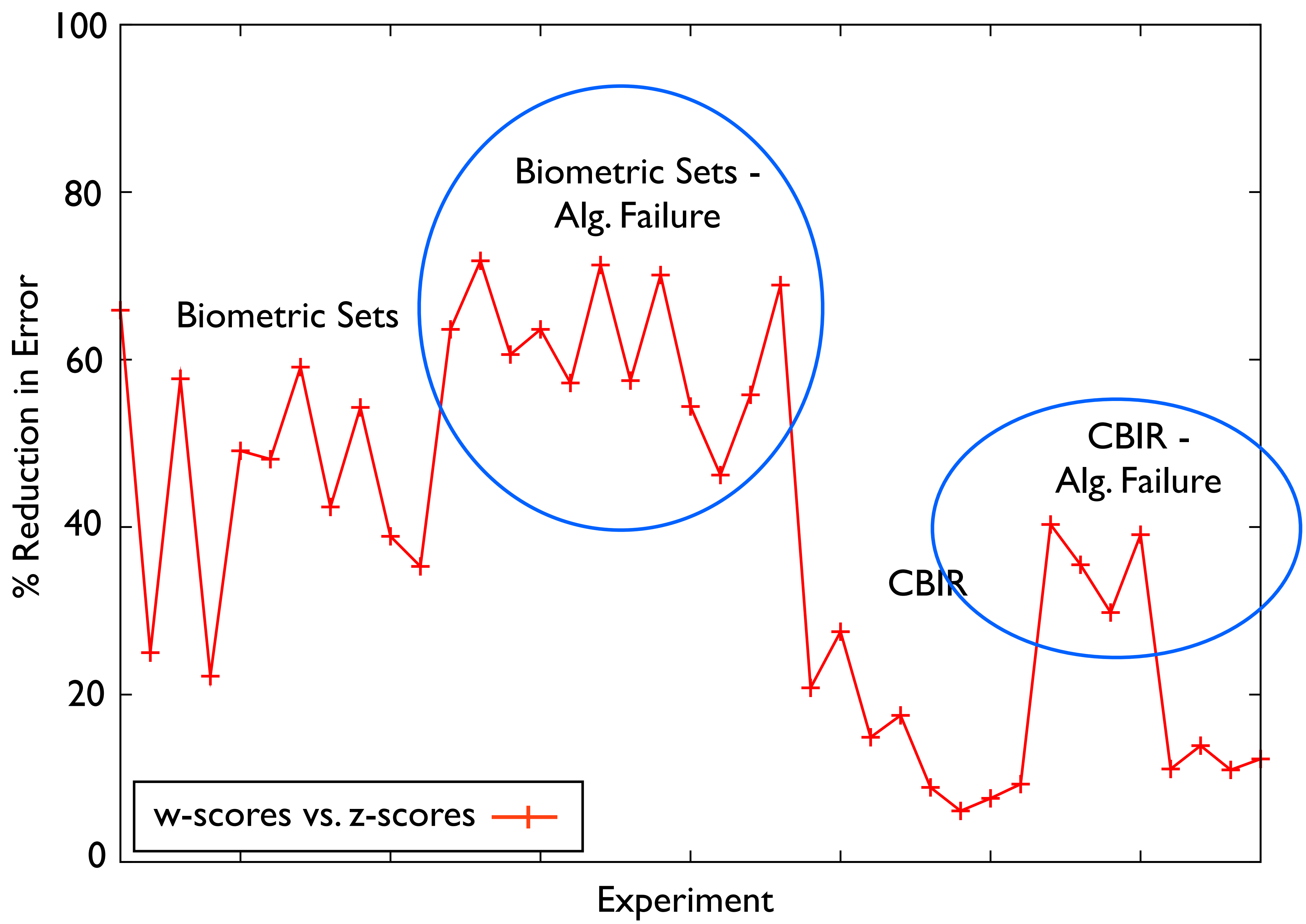
3. **While** $k < m$ **do**

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

5. $\quad k = k + 1$

6. **end while**

Error Reduction: Failing vs. Succeeding Algorithm



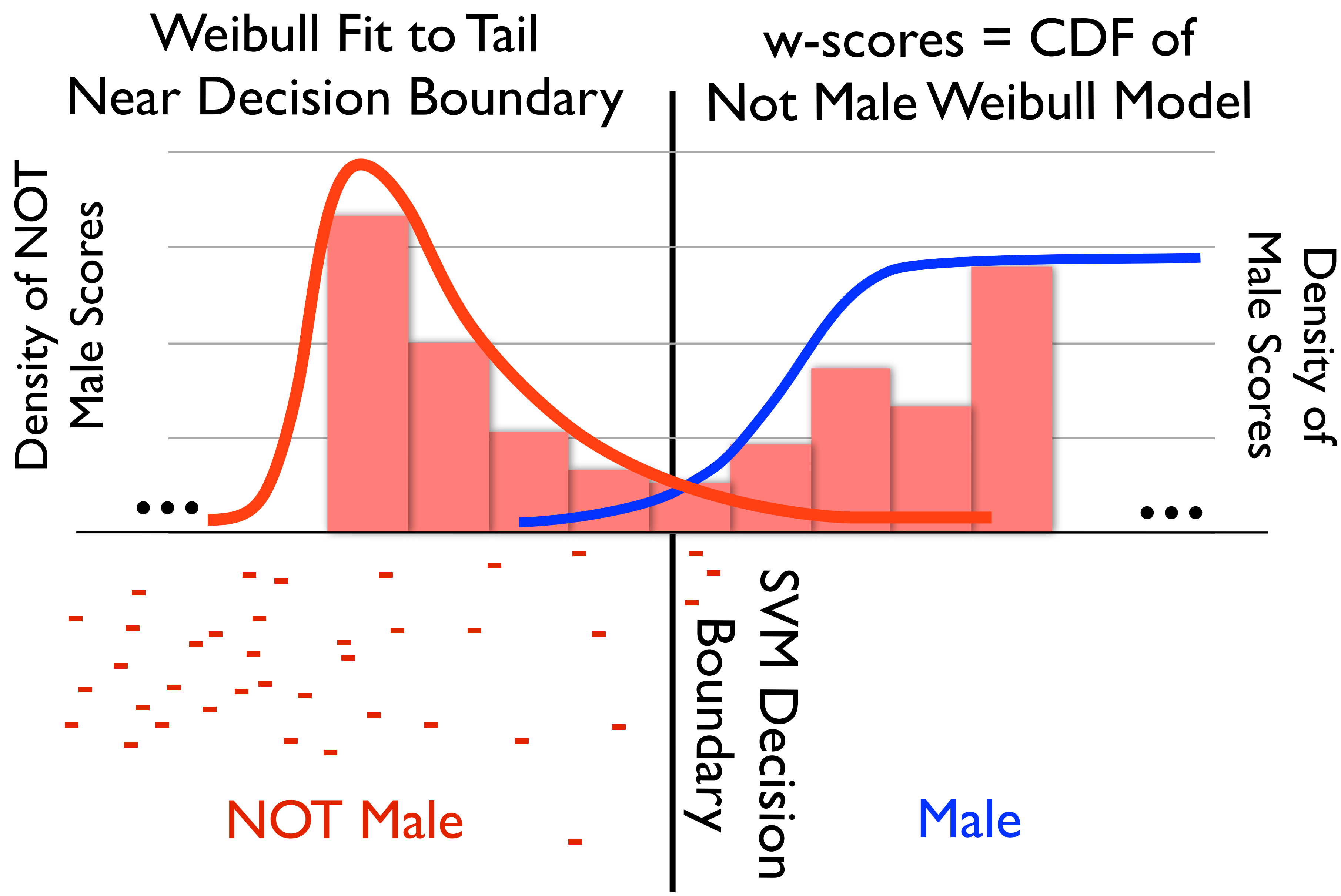
Multi-Attribute Spaces

- Let $P(L(j)|I)$, $j = 1...N$, be the probability that humans would assign label $L(j)$ to a given image I
- Let $\mathrm{A}_j(I)$ be attribute classifiers that map images to real-valued scores
- Let $\mathrm{E}(A_j) \equiv |A_j(I) - P(L(j)|I)|$ be the expected labeling error in A_j

Multi-Attribute Spaces

- Definition 1. A continuous function $A_j : I \mapsto [0,1]$ is called a well normalized attribute function when $E(A_j(I)) \leq \varepsilon$ with a probability of at least $1 - \delta$

- Definition 2. A multi-attribute space $M : I \mapsto [0,1]^N$ is a product space formed from well normalized attribute functions, $M(I) = A_1(I) \times A_2(I) \times ... \times A_N(I)$



Weibull Fit to Tail
Near Decision Boundary
w-scores = CDF of
Not Male Weibull Model
Density of NOT
Male Scores
Density of
Male Scores
SVM Decision
Boundary
NOT Male
Male

Fusion for Multi-Attribute Search

Solve the following problem:

maximize over I $\qquad s^q = \| A_j(I) \|_1$

subject to $\qquad A_j(I) = \mathrm{CDF}(s_j(I); W_j);$

for $\forall\, j \in J$ satisfying $\qquad 0 \leq \alpha_j \leq A_j(I) \leq \beta_j \leq 1;$

Goal: find the images that maximize the L_1 norm of estimated probabilities for each attribute that also satisfy the constraints α_j and β_j

Multi-Attribute Search

"Indian Females"

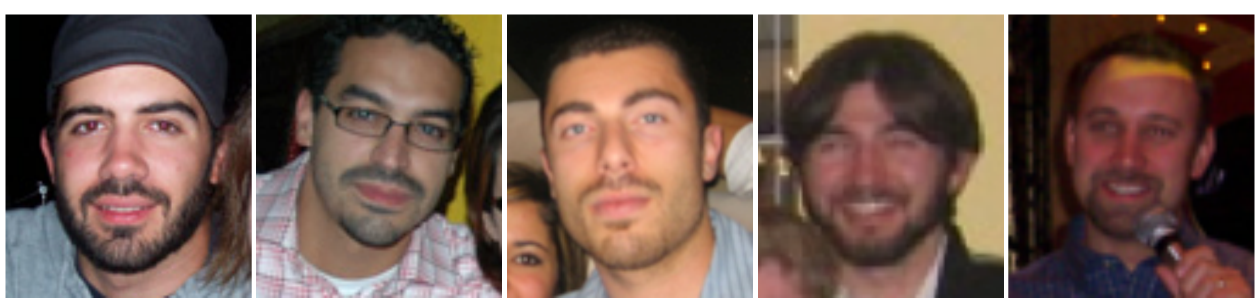
Our Approach

Comparison with the approach presented by Kumar et al. in T-PAMI 2011

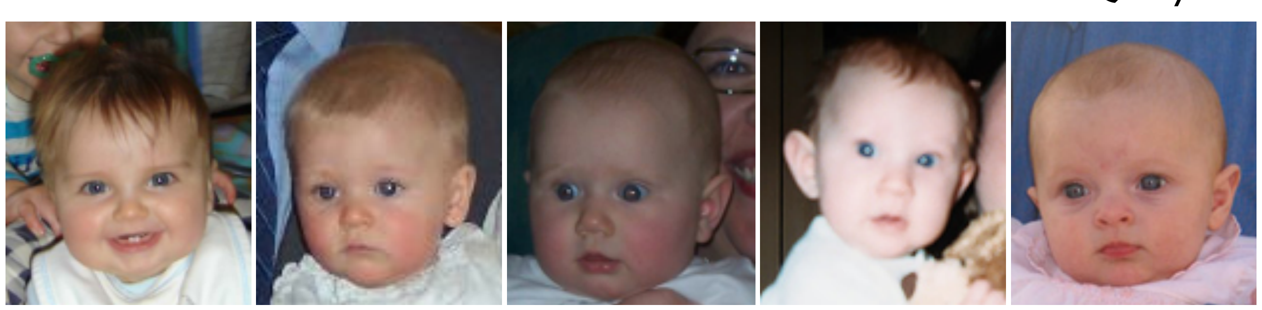
Kumar et al. 2011 | **Our Multi-Attribute Space Approach**

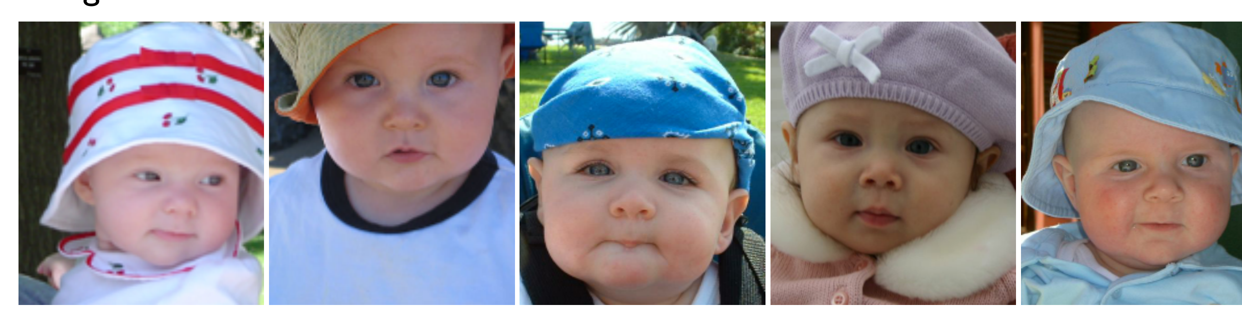
Query: Women with Pale Skin

Query: Chubby Indian Men with Mustache



Query: White Babies Wearing Hats



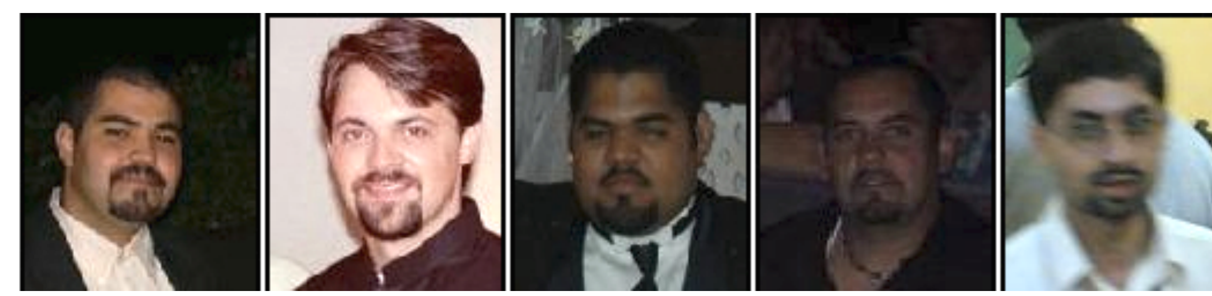


Comparison with the approach presented by Kumar et al. in T-PAMI 2011

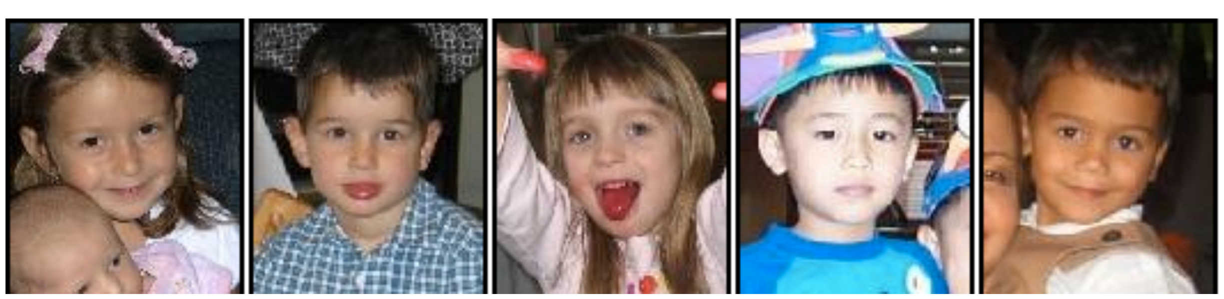
Kumar et al. 2011 | **Our Multi-Attribute Space Approach**

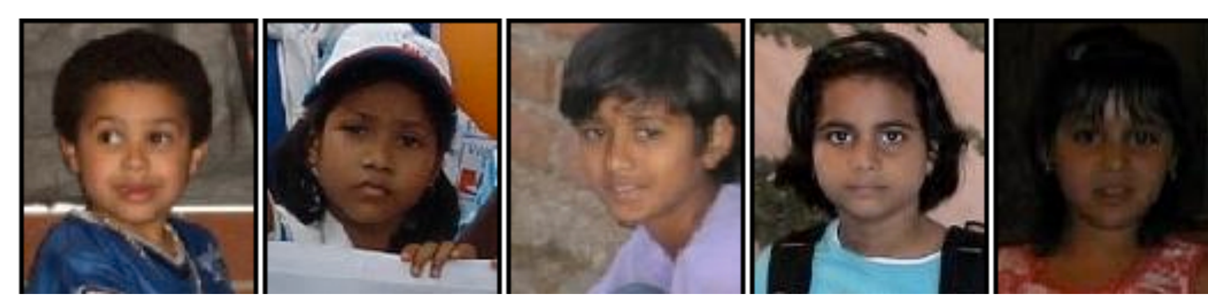
Query: Women with Curly Hair

Query: Men with Black Hair and Goatee



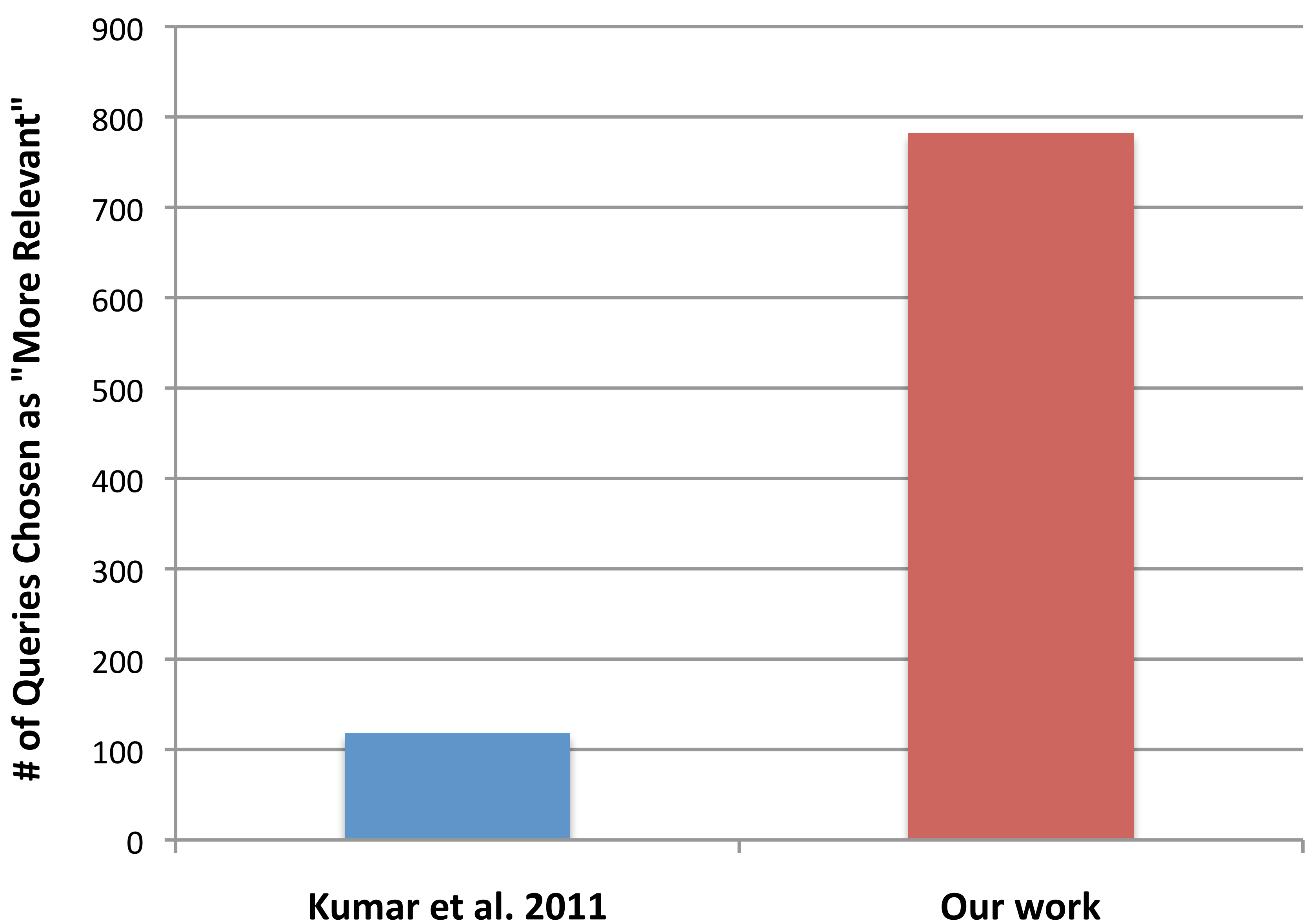
Query: Indian Kids with Round Face





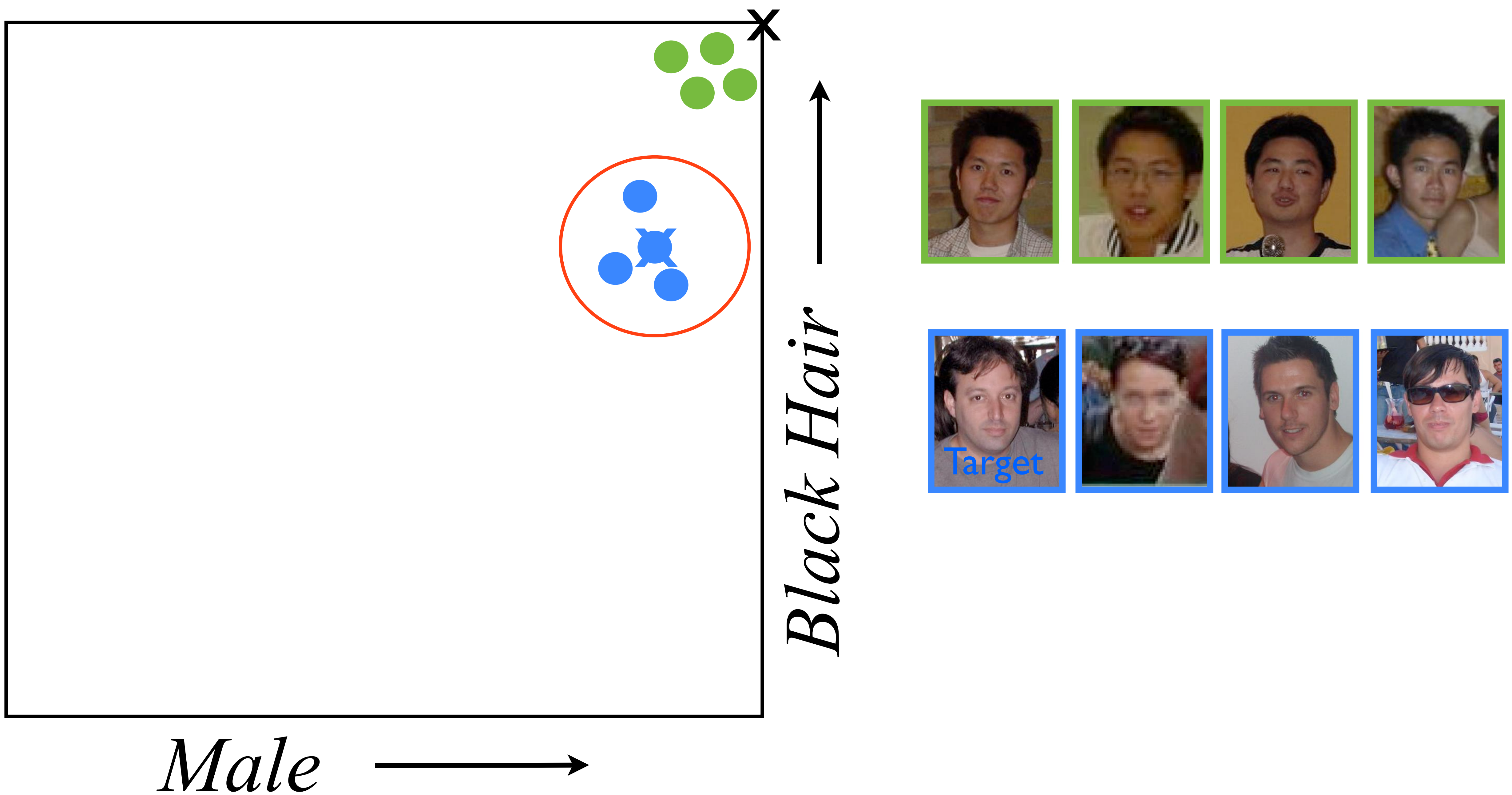
Comparison with the approach presented by Kumar et al. in T-PAMI 2011

For 900 comparison tests, our approach was selected as "more relevant" **86.9%** of the time

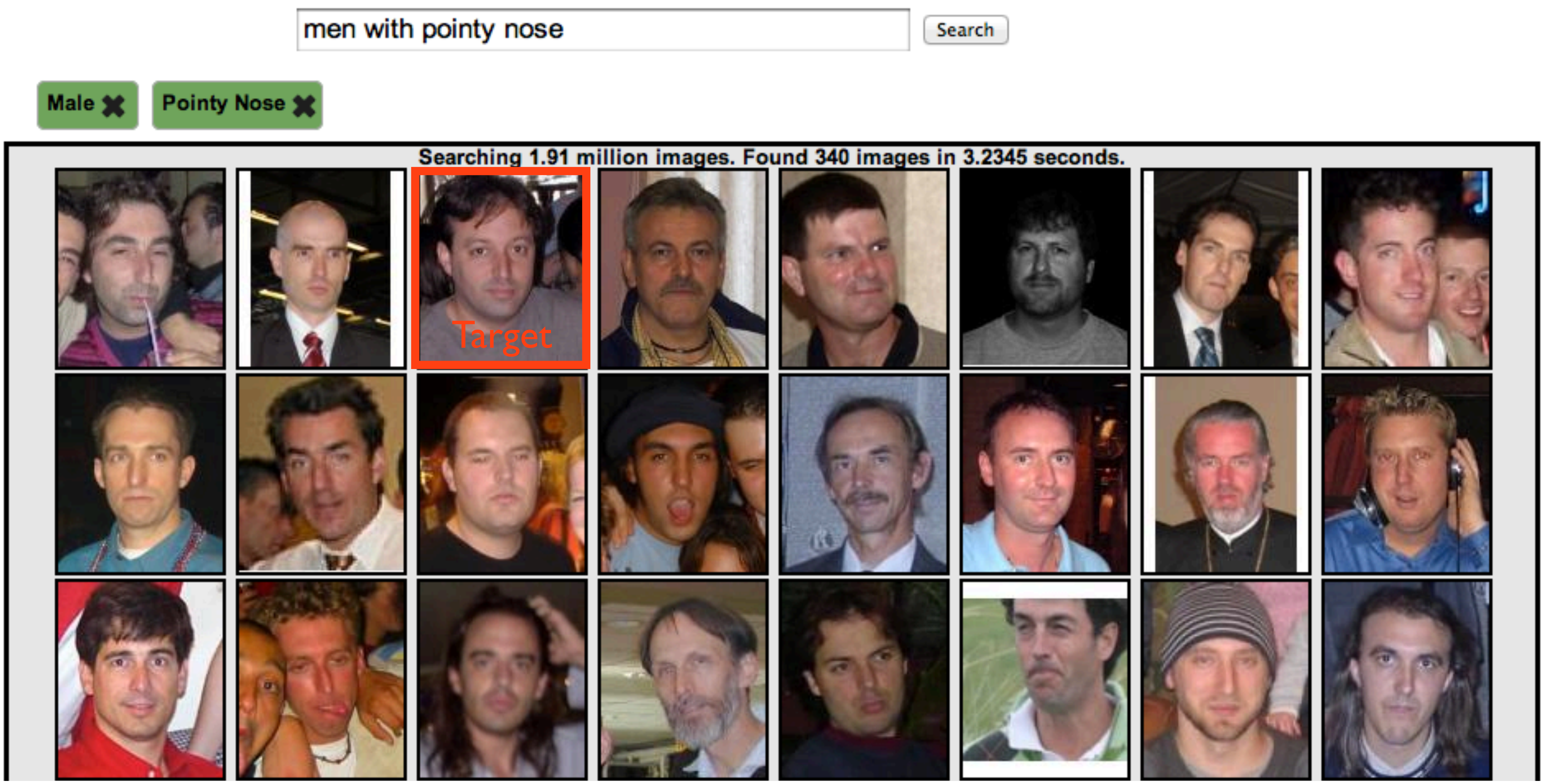


Similar Attribute Search

For finer grained search, we are interested in candidates outside of just the top results with the highest scores



A new way to search: similarity search based on target attributes from a particular image

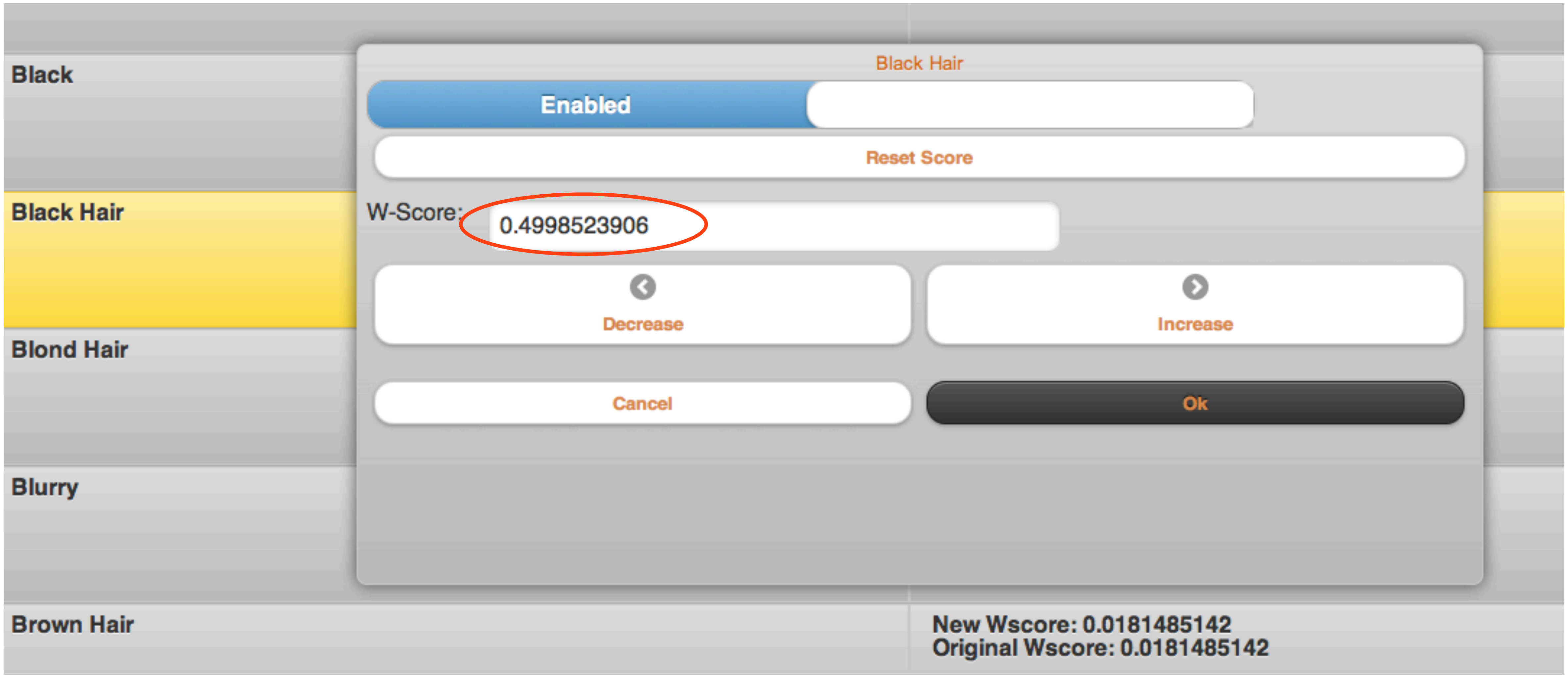


Target Attribute Details

Query	
Attribute	**Weight**
Male	0.8122
Pointy Nose	0.8687
Perfect:	**1.6809**

W-Scores for query			
Attribute	**Calculation Steps**		**W-Score**
Male	Initial w-score	0.9978451603	0.810449839196
	Re-weighted w-score	0.810449839196	
Pointy Nose	Initial w-score	0.9939414012	0.863436895222
	Re-weighted w-score	0.863436895222	
Total:			**1.67388673442**
Rank:			**99.58%**

Additional target attributes from the chosen image can be added to Refine the Query:

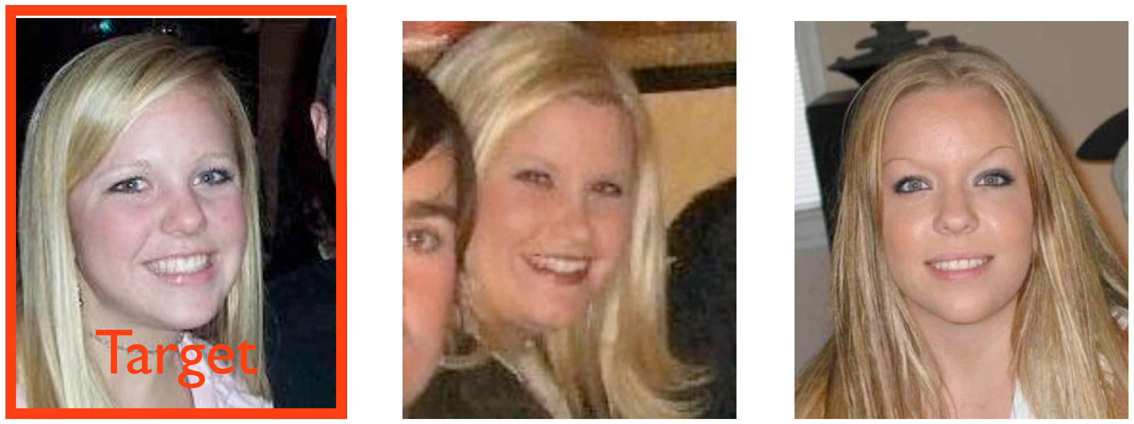


Similar Attribute Search Results

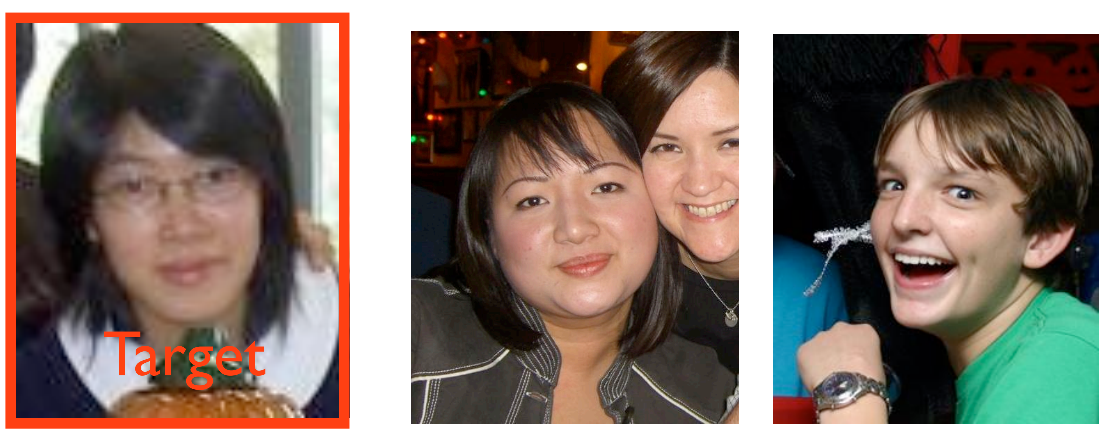
Query: Men with a Pointy Nose and Black Hair like the targets in the selected image

Similar Attribute Search Results

Query: Blonde hair like the target in the selected image



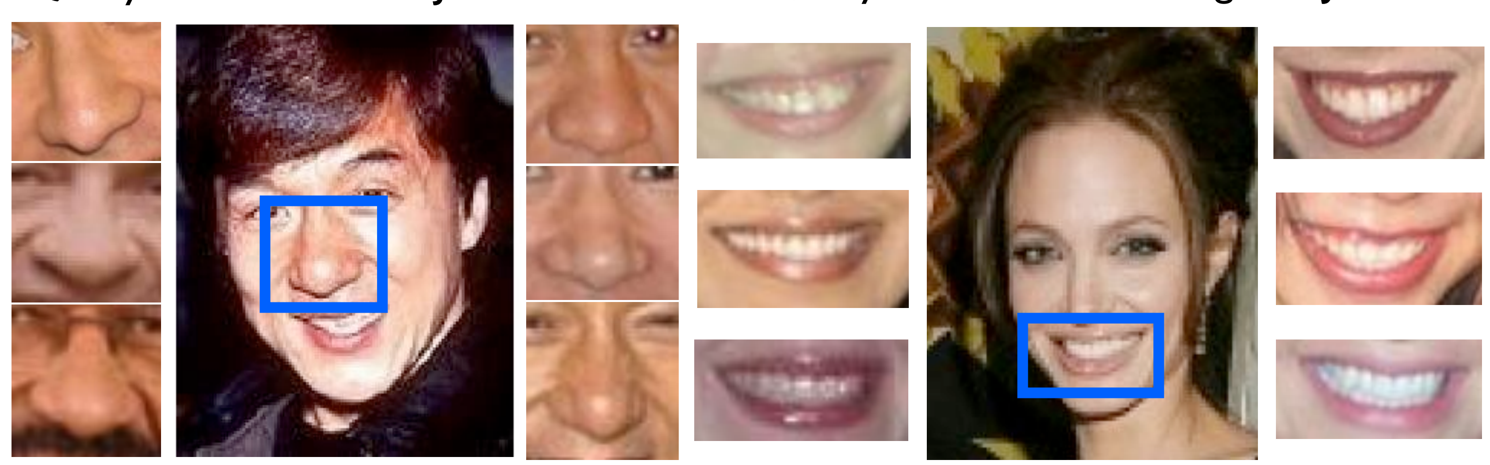
Query: Black Hair and Bangs like the targets in the selected image



Query: Beard, Pointy Nose and Pale Skin like the targets in the selected image

Queries can be mapped to specific names:

Query: Nose Most like Jackie Chan's Query: Smile Most like Angelina Jolie's

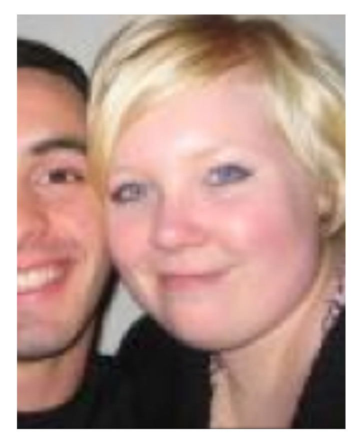


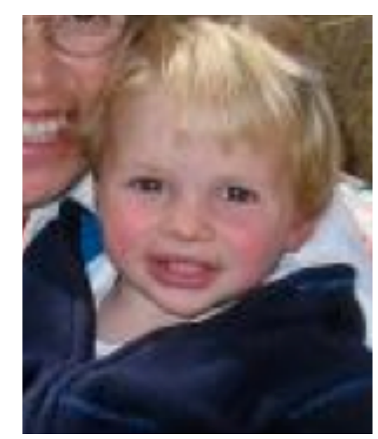
Two Approaches to Results Ordering

Target

Query: Rosy Cheeks
& Blonde Hair Most
Like this image

Ordering Based on Distance
Measured from Query Attributes





Ordering Based on Distance Measured from
Query Attributes + Other Contextual Attributes

Ordering Based on Distance From Target Attributes for Query Attributes

Query: Blonde Hair and Rosy Cheeks like Selected Image

Statistically significantly better than an ordering not consistent with human ordering, with a p value < 0.01

Ordering Based on Distance Measured from Query Attributes + Other Contextual Attributes

Query: Blonde Hair and Rosy Cheeks like Selected Image

Statistically significantly better than an ordering not consistent with human ordering

Statistically significantly better than an ordering based just on query attributes

Ordering Based on Distance From Target Attributes for Query Attributes

Query: Chubby Face and Round Face like selected Image

Statistically significantly consistent with human ordering

Ordering Based on Distance Measured from Query Attributes + Other Contextual Attributes

Query: Chubby Face and Round Face like selected Image

Statistically significantly better than an ordering not consistent with human ordering

Statistically NOT significantly better than an ordering based just on query attributes

Try this out

- The search engine: http://mughunt.securics.com
- The attribute service: http://afs.automaticfacesystems.com/
- The Meta-Recognition library: http://www.metarecognition.com/

Acknowledgements

- Neeraj Kumar, UW
- Anderson Rocha, UNICAMP
- Peter Belhumeur, Columbia University
- Terry Boult, UCCS
- Ross Micheals, NIST

Questions?