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 2007¹
  - Describe objects by their *attributes*

  Has Horn
  Has Leg
  Has Head
  Has Wool

- Kumar et al. T-PAMI 2011²
  - Describe faces by their *attributes*

  Has Hat
  Has Beard
  Has African Ethnicity
  Has Round Nose


Top Image by Cliff Hall “Mountain Goat” BY-NC-ND http://www.flickr.com/photos/cliffhall/303337039/in/photostream/

Bottom Image by Enrico Fuente “Ghostface Killah” BY-NC-ND http://www.flickr.com/photos/okobojierik/5156583220/in/photostream/
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:
      \[
      \mathcal{F}(I) = \{f_1(I), \ldots, f_k(I)\}
      \]
    - 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 \textit{w-scores}, which are estimated probabilities of an attribute NOT being drawn from the class opposite to it.

A \textbf{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.

What is this?

- Teapot
  - Red teapot
- Frog
  - Frog on corn leaf
- Girl
  - Lovely little girl:

Image by Olivier Ffrench “Quiet brown frog” BY http://www.offrench.net/
Image by Joi Ito “Frog on corn leaf” BY http://www.fotopedia.com/users/joi/
Image by BirdCantFly “Lovely little girl:)” BY http://www.flickr.com/photos/birdcantfly/
Image by fraise “Red teapot” BY http://www.flickr.com/photos/fraise/
Data Fusion

- A single algorithm is not a complete solution for a recognition task
- Combine information across algorithms, classifiers, or sensors

  - Decision fusion
  - Score level normalization & fusion

Do this is a robust manner...

Met​a-Rec​ogni​tion

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

- Recognition System
- Meta-Recognition System (Generic Predictor)
- Post. Recognition Score Data
- Prediction
- Success/

- Request Operator Interaction
- Perform Fusion
- Ignore Data
- Acquire More Data
- etc.

Re-Start

From Meta-Cognition to Recognition

• Inspiration: *Meta-Cognition* Study

  - “knowing about knowing”

  - Example: If a student has more trouble learning history than math, she “knows” something about her learning ability and can take corrective action

---

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\textsuperscript{1}: three different quality assessment algorithms lacked correlation
  - Face\textsuperscript{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

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
  - 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

---

Recognition Systems

- Overall Non-Match Distribution
- True Rejection
- Overall Match Distribution
- False Recognition
- True Recognition
- Post-Recognition Non-Match Scores Histogram
- Post-Recognition Match Score
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:

\[
c^* = \arg\max_{c} \Pr(p_0 = p_c)
\]

subject to \(\Pr(p_0 = p_{c^*}) \geq 1 - \delta\), for a given confidence threshold \(\delta\). 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.

---

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 $F(p)$ 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) : $\forall x \in S(K), x \in F(p)$
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

Overall Distribution of Scores

Distribution’s tail

Match

Extrema

Extreme Value Theory

Tail Analysis

Best of Portfolio Matches

Portfolios of Gallery Scores

Scores

Frequency

Portfolios

Scores

Extrema
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\(^2\).

---

The Weibull Distribution

The sampling of the top-$n$ scores always results in an EVT distribution, and is \textit{Weibull} if the data are bounded$^1$.

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

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

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** $\text{Inv}(W_S(s_1)) > \delta$ **do**

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

6. **end if**

$\delta$ 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

<table>
<thead>
<tr>
<th></th>
<th>Conventional Explanation</th>
<th>Prediction</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>False Accept</td>
<td>Success</td>
<td>O</td>
</tr>
<tr>
<td>Case 2</td>
<td>False Reject</td>
<td>Failure</td>
<td>O</td>
</tr>
<tr>
<td>Case 3</td>
<td>True Accept</td>
<td>Success</td>
<td>P</td>
</tr>
<tr>
<td>Case 4</td>
<td>True Reject</td>
<td>Failure</td>
<td>P</td>
</tr>
</tbody>
</table>

**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

Face Recognition

Points approaching the lower left corner minimize both errors

Statistical Meta-Recognition
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:

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

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, $s_1, \ldots, s_n \in S$;

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

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

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

5. $k = k + 1$

6. **end while**
Error Reduction: Failing vs. Succeeding Algorithm

% Reduction in Error

Biometric Sets

CBIR

w-scores vs. z-scores

Experiment
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 $A_j(I)$ be attribute classifiers that map images to real-valued scores.

- Let $E(A_j) ≡ |A_j(I) - P(L(j) | I)|$ 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

SVM Decision Boundary

Density of Male Scores

NOT Male

Male
Fusion for Multi-Attribute Search

Solve the following problem:

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

subject to  \[ A_j(I) = \text{CDF}(s_j(I); W_j); \]

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

Goal: find the images that maximize the $L_1$ norm of estimated probabilities for each attribute that also satisfy the constraints $\alpha_j$ and $\beta_j$
Multi-Attribute Search

“Indian Females”

Indian → Male

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

Kumar et al. 2011

Query: Women with Pale Skin

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.8122</td>
</tr>
<tr>
<td>Pointy Nose</td>
<td>0.8887</td>
</tr>
<tr>
<td>Perfect</td>
<td>1.6809</td>
</tr>
</tbody>
</table>

W-Scores for query

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Calculation Steps</th>
<th>W-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Initial w-score 0.9973451603</td>
<td>0.810449839196</td>
</tr>
<tr>
<td></td>
<td>Re-weighted w-score 0.810449839196</td>
<td></td>
</tr>
<tr>
<td>Pointy Nose</td>
<td>Initial w-score 0.9939414012</td>
<td>0.863436895222</td>
</tr>
<tr>
<td></td>
<td>Re-weighted w-score 0.863436895222</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1.67389673442</td>
</tr>
<tr>
<td>Rank</td>
<td></td>
<td>99.58%</td>
</tr>
</tbody>
</table>
Additional target attributes from the chosen image can be added to Refine the Query:

- Black Hair: 0.4998523906
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

Ordering Based on Distance Measured from Query Attributes

Target

Query: Rosy Cheeks & Blonde Hair Most Like this image

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?