Meta-Recognition, Machine Learning and the Open Set Problem

Walter J. Scheirer
Director of Research & Development, Securics, Inc.
Assistant Professor Adjoint at the University of Colorado at Colorado Springs
What is recognition in computer vision?

- Compare an object to a known set of classes, producing a similarity measure to each.
Why is recognition hard?

The same object can cast an infinite number of different images onto the retina (humans) or an innumerable number of images on a sensor (machine).


Image by Michele Catania “Eye” BY http://www.flickr.com/photos/cataniamichele/
Why is recognition hard?
Why is recognition hard?

What strategies do we have to approach this problem?

- Multiple-View Geometry
- 3D Modeling
- Invariant Feature Descriptors
- Data Fusion
- Machine Learning
Data Fusion

• A single algorithm is not a complete solution for a recognition task

• Combine information across algorithms and sensors¹
  - Decision fusion
  - Score level normalization & fusion

Do this is a robust manner...

**Meta-Recognition**

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

- **Recognition System**
  - **Post. Recognition Score Data**
  - **Monitoring**
  - **Meta-Recognition System** (Generic Predictor)
  - **Prediction**
    - **Success?**
    - **Done**
    - **Failure**

- **Control**
  - **Re-Start**
  - **Request Operator Interaction**
  - **Perform Fusion**
  - **Ignore Data**
  - **Acquire More Data**
  - **etc.**

---

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

J. Flavell and H. Wellman, “Metamemory,” in Perspectives on the Development of Memory and Cognition, 1988, pp. 3-33
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?

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...

Apparent quality is not always tied to rank.
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

Overall Match Distribution

Post-Recognition Non-Match Scores Histogram

Post-Recognition Match Score

f(x)

x

True Recognition

False Recognition

False Rejection

True Rejection

$t_0$
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:

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

subject to \(\Pr(p_0 = pc^*) \geq 1 - \delta\), 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

Portfolios of Gallery Scores

Best of Portfolio Matches

Distribution’s tail

Match

Extrema

Extreme Value Theory

Tail Analysis
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 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** $\text{Inv}(W_S(s_1)) > \delta$ **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>Case</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

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

Meta-Recognition Miss Detection Rate

\[
MRFAR = \frac{|\text{Case 2}|}{|\text{Case 2}| + |\text{Case 3}|}
\]
Comparison with Basic Thresholding over Original and T-norm Scores

Face Recognition

Points approaching the lower left corner minimize both errors

Graphical representation showing the comparison of different thresholding methods in the context of face recognition. The axes represent different error metrics, and the curves illustrate how different techniques perform under varying conditions.
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:

\[
CDF(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}
\]

We call this a \textit{w-score}
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

Experimt

w-scores vs. z-scores
Let’s take a step back and consider machine learning for recognition...

• Large-scale learning is a major recent innovation in computer vision
  - Feed lots of features to a learning algorithm, and let it find correlation

• How should we approach the multi-class problem\(^1\) for general object recognition?
  - Is it a series of binary classifications?
  - Should it be performed by detection?
  - What if the classes are ill-sampled, not sampled at all, or undefined?

---

Closed Set Recognition

- How well are we really doing on recognition tasks?
- The problem we’d like to solve: scene understanding given an image never seen before
- The problem data sets solve: given a set of known classes, and corresponding ‘+’ and ‘-’ labels, distinguish between these classes
  - Caltech 101 & 256
  - LabelMe
  - ImageNet
- Training and Testing on the same data

Closed Set Recognition

Open Set Recognition

• There are classes not seen in training that occur in testing

• Suppose the “other” classes are known
  ▶ we generally cannot have enough positive samples to balance the negative samples

“All positive examples are alike; each negative example is negative in its own way”

Open Set Recognition

Formalization of Open Set Recognition Problem

• A class is a distribution $\mathcal{P}$
• A sample $V$ is labeled $L = +1$ if it belongs to the class to be recognized and $L = -1$ for any other class
• Training samples from $\mathcal{P}$: $\hat{V} = \{v_1, \ldots, v_m\}$
• Training samples from other known classes $\mathcal{K}$: $\hat{K} = \{k_1, \ldots, k_n\}$
• The larger universe of unknown negative classes $\mathcal{U}$
• Test data: $\{t_1, \ldots, t_z\}$, $t_i \in \mathcal{P} \cup \mathcal{K} \cup \mathcal{U}$
• A measurable recognition function $f$ for a class $\mathcal{P}$

Recognition Risk: $R(f) = \mathbb{E}(\text{sign}(f(V)) \neq L)$
Formalization of Open Set Recognition Problem

• A few notes on Risk
  
  - Ensure that the risk of a false positive (over generalization) is proportional to the volume of space which is labeled positive
  
  - Ensure that over specialization occurs if we define the region too narrowly around the training data
  
  - Good solutions to the open set recognition problem require minimizing the volume of space representing the learned recognition function \( f \)
    
    ▶ Outside the support of positive samples
Formalization of Open Set Recognition Problem

- We also need to optimize a data error measure:

\[ D(f(v_i); f(k_j)); (v_i \in \hat{V}, k_j \in \hat{K}) \]

\( D \) could be: inverse F-measure over the training data, inverse training precision for a fixed training recall, inverse training recall for a fixed training precision...

Goal: balance the risk with the data error measure, all while being subject to hard constraints from the positive training data and/or negative training data.
Formalization of Open Set Recognition Problem

The Open Set Recognition Problem

Using training data with positive samples, and other known class samples, and a data error measure, find a measurable recognition function $f$, where $f(x) > 0$ implies positive recognition, and $f$ is defined by:

$$\text{argmin} \{R(f) + \lambda_r D(f(v_i); f(k_j))\}$$

subject to

$$m\alpha \leq \sum_{i=1}^{m} \phi(f(v_i)) \quad \text{and} \quad n\beta \geq \sum_{j=1}^{n} \phi(f(k_j))$$

where $\lambda_r$ specifies the regularization tradeoff between risk and data, where $\alpha \geq 0$ and $\beta \geq 0$ allow a prescribed limit on true positive and/or false positive rates, and $\Phi$ is a given loss function.
The trouble with binary classification
The trouble with 1-vs-All classification
One Solution: 1-class SVM

- Formulation by Schölkopf et al.¹

  - Origin defined by the kernel function serves as the only member of a “second class”
  
  - Find the best margin with respect to the origin
  
  - The resulting function \( f \) takes the values
    
    ‣ +1 in a region capturing most of the training data points
    
    ‣ -1 elsewhere

---

One Solution: 1-class SVM

To separate the training data from the origin, the algorithm solves the following quadratic programming problem for $w$ and $\rho$ to learn $f$:

$$
\min \frac{1}{2} \| w \|^2 + \frac{1}{\nu m} \sum_{i=1}^{l} \xi_i - \rho
$$

subject to

$$(w \cdot \Psi(x_i)) \geq \rho - \xi_i \quad i = 1, 2, \ldots, m \quad \xi_i \geq 0$$

The kernel function $\Psi$ impacts density estimation and smoothness. The regularization parameter $\nu \in (0, 1]$ controls the trade-off between training classification accuracy and the smoothness term $\| w \|$, and also impacts the number of support vectors.
I-Class SVM

Generalization

1-class SVM $\nu = 0.04 \gamma = 0.25$

Specialization

1-class SVM $\nu = 0.04 \gamma = 8$
Start with a 1-class SVM

- Base1-vs-Set Near Plane A
- Base1-vs-Set Far Plane Ω
- Org
Generalization

Generaliized 1-vs-Set Far Plane $\Omega$

Base 1-vs-Set Near Plane

Org

Generalized 1-vs-Set Far Plane $\Omega$

Base 1-vs-Set Near Plane

Org
Specialization

Specialized 1-vs-Set Far Plane

Org

Specialized 1-vs-Set Near Plane A

Base 1-vs-Set Near Plane A

Base 1-vs-Set Far Plane Ω
Where is this work heading?

• The 1-vs-Set Machine as an initial solution for open set recognition

• New classes of learning algorithms to specifically address the open set problem

• Application Area: Computational Linguistics
  ▸ The recognition problem occurs here too

Taking literary theory into practice!
Sometimes confusion is a good thing...
Questions?