Statistical Models for Assessing How Features of the Human Face Affect Recognition

Geof Givens
Statistics Department
Colorado State University
Fort Collins, CO

Ross Beveridge
Computer Science Department
Colorado State University
Fort Collins, CO

Bruce Draper

Patrick Grother
Information Technology Lab.
NIST
Gaithersburg, MD

Jonathon Phillips

Subject/Image Covariates

Race

Gender

Age

Eyes

Facial Hair

Glasses

Bangs

Mouth

Smiling?
Subject Image Data

- 1,072 Human Subjects from the FERET Data
- 2,144 FERET Images
- Exactly 2 images per subject, taken on same day
Collecting the Covariates
## Our Subject Covariate Factors

### FERET Subject/Image Covariates

<table>
<thead>
<tr>
<th>Fixed Per Subject</th>
<th>Fixed Per Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Bangs</td>
</tr>
<tr>
<td>Gender</td>
<td>Expression</td>
</tr>
<tr>
<td>Race</td>
<td>Eyes</td>
</tr>
<tr>
<td>Skin</td>
<td>Facial Hair</td>
</tr>
<tr>
<td></td>
<td>Makeup</td>
</tr>
<tr>
<td></td>
<td>Mouth</td>
</tr>
<tr>
<td></td>
<td>Glasses</td>
</tr>
<tr>
<td>Young</td>
<td>No</td>
</tr>
<tr>
<td>Old</td>
<td>Yes</td>
</tr>
<tr>
<td>Male</td>
<td>Neutral</td>
</tr>
<tr>
<td>Female</td>
<td>Other</td>
</tr>
<tr>
<td>White</td>
<td>Open</td>
</tr>
<tr>
<td>Black</td>
<td>Other</td>
</tr>
<tr>
<td>Asian</td>
<td>No</td>
</tr>
<tr>
<td>Other</td>
<td>Yes</td>
</tr>
<tr>
<td>Clear</td>
<td>Neutral</td>
</tr>
<tr>
<td>Other</td>
<td>Open</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Closed</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
Studies can also include covariates for...

- Algorithm type (see later slides)
- Algorithm configuration (tunings, aspects of training, etc.)
- Imaging characteristics (darkness, lighting angle, eye calibration, etc.)
- Time difference between images
Advantages to incorporating many covariates simultaneously...

- **Marginal comparison of men vs. women can identify which group is easier to recognize, but not why.**
- **Covariates can be correlated, so their individual effects are hard to isolate.** E.g., if men are easier to recognize, would it be because of their gender, frequent facial hair, preference for glasses instead of contacts, or what?
- **Covariates can interact.** E.g., do glasses ease recognition for all algorithms equally?
We will study subject covariate effects for three recognition algorithms: First Algorithm: PCA

Training

Training images

Eigenspace

Testing

Distance Matrix

Whitened Cosine

PCA space projection

Second Algorithm (IIDC): Intrapersonal Image Difference Classifier

Training

- Classify difference image as either:
  - Intrapersonal from same subject
  - Extrapersonal from different subjects

Intrapersonal Example

Extrapersonal Example

Second Algorithm (IIDC): Intrapersonal Image Difference Classifier

Testing

Extrapersonal PCA Subspace

Intrapersonal PCA Subspace

For each pair of probe and gallery images
- Compute difference
- Rank by intraperpersonal class probability.

Distance Matrix
Third Algorithm (EBGM): Elastic Bunch Graph Matching

Training
Hand Select Landmarks on Training Subjects

Extract Gabor Jets
One jet per Landmark.

Build Bunch Graph
One node per landmark.
Holds jets for model subjects.

Typical number: 25 Landmarks, 70 model subjects.

Okada et. al., The Bochum/USC Face Recognition System and How it Faired in the FERET Phase III Test, in Face Recognition: From Theory to Applications
Fourth Algorithm (EBGM): Elastic Bunch Graph Matching

Testing
Localize Landmarks

Build Face Graphs

Compare Face Graphs

Distance Matrix
Performance Data

- Obtain a similarity score (distance) by randomizing images into probes and gallery
- FERET data still valuable: far more individuals than other datasets like Notre Dame data
Response Variables

• Similarity score (distance) between each image and its true match. Continuous, but difficult to normalize, especially across multiple algorithms.

• Rank-one recognition. Binary: is the similarity score larger than scores between the probe and all gallery images of other people? 1=yes 0=no.
Linear Model

Response = Similarity score

$Y_i = \text{Similarity score for image pair } i \text{ (normalized)}$

$X_i = \text{Algorithm \& human covariate factors for image pair } i.$

$\beta = \text{Parameters quantifying factor effects.}$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \varepsilon_i$$

with $\varepsilon_i \sim \text{iid Normal}(0, \sigma^2)$
Generalized Linear Model

Response = rank-one recognition

\( Y_i = \) Was the \( i \)th image pair matched at rank 1?  
(i.e. \( Y_i = 1 \) if \( R_i = 1 \) and otherwise \( Y_i = 0 \))

\( X_i = \) Algorithm & human covariate factors for image pair \( i \).

\( \beta = \) Parameters quantifying factor effects.

\( Yi \mid Xi \sim f(y \mid \mu_{Yi \mid Xi}) \) independently, where

\[
g(\mu_{Yi \mid Xi}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \varepsilon_i
\]

Now:

\( f(\mu_{Yi \mid Xi}) = \text{Bernoulli}(\mu_{Yi \mid Xi}) \),

\( g(z) = \log(z/(1-z)) \),

\( \varepsilon_i \sim \text{iid Normal}(0, \sigma^2) \)
Some people are simply harder than others

- When we have many images per subject...
- Subjects are viewed as randomly drawn from some population, and their individual effects are not interesting but their variation and within-person correlation must be accounted for.
- This effectively removes person-to-person variation from the analysis, providing better resolution for detecting effects not attributable to natural variation among people.
- Model has correlated covariance structure, since multiple responses from same person are dependent.
Generalized Linear Mixed Model

Response = rank-one recognition

$Y_i = \text{Rank-one recognition}$

$X_i = \text{Algorithm & image fixed covariate factors}$

$Z_i = \text{Indicators for random factors, say } Z_{i1} \ldots Z_{ir}.$

$\beta = \text{Parameters quantifying fixed factor effects.}$

$\gamma = \text{Parameters quantifying random effects.}$

$Y_i \mid X_i \sim f(y \mid \mu_{Y_i \mid X_i})$ independently, where

$g(\mu_{Y_i \mid X_i}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \gamma_1 Z_{i1} + \ldots + \gamma_r Z_{ir} + \varepsilon_i$

where:

$f(\mu_{Y_i \mid X_i}) = \text{Bernoulli}(\mu_{Y_i \mid X_i})$

$g(z) = \log \left( \frac{z}{1-z} \right),$ 

$\gamma_i \sim \text{iid Normal}(0, \tau^2)$

$\varepsilon_i \sim \text{iid Normal}(0, \sigma^2)$
GLM with Three Algorithms

<table>
<thead>
<tr>
<th>Feature</th>
<th>EBGM</th>
<th>IIDC</th>
<th>PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Glasses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes Change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Eyes Not Open</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expression Changes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Other Expression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Other Skin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Other Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Asian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Black</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth Changes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Other Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facial Hair Changes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Facial Hair</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangs Change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Bangs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject Old</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Change to Baseline Predicted Pr[Rank 1 Recognition]
Results: Gender Effect???

From FRVT 2002 Evaluation Report

These results are based upon simple partition of probe set!

Fig. 17. CMC showing the difference in performance between male and female probes. For each rank, male minus female performance is plotted. The CMC shows that the identification rate of males is better than females.
No Significant Gender Effect Found After Controlling for Confounding Factors

GLM: No significant change to baseline predicted rank-one recog. rate

Female

HARDER  EASIER

-0.02  0.0  0.02

EBGM  IIDC  PCA

...after adjusting for potentially confounding factors!
Significant Effect for Age, with no Algorithm Interaction

GLM: Significant increase in rank-one recog. rate; all algorithms affected similarly
Glasses Affect Different Algorithms Differently

GLM: Glasses improve rank-one recog rate for PCA; reduce it for IIDC; and have little effect for EBGM. **Significant interaction.**

Result not shown here: Removing/adding glasses between probe and gallery hinders recognition!
Another significant interaction: eyes and algorithms

GLM: Closing your eyes significantly impedes EBGM compared to the other two algorithms.

Speculation: EBGM suffers when eyes cannot be reliably located for landmarks.
Concluding Thought

- Algorithm evaluation is an excellent arena for these types of methods because simulation experiments can be designed in the computer.

- Randomization, balance, other criteria of statistical design are usually easy to achieve.

- Aside from coding covariates, little human effort is required to produce vast datasets.

- Usually in statistics, we must craft a model to cope with the data available. Here, we may choose a model to best answer our questions of interest, then design the experiment and collect the data required for that model.
Thank You

More info: www.cs.colostate.edu/evalfacerec/