Assessing the Effect of Crossing Databases on Global and Local Approaches for Face Gender Classification

Y. Andreu, R.A. Mollineda and P. García-Sevilla

Universitat Jaume I

Problem

Which approach is more suitable for face gender classification when acquisition and demographic conditions of the images vary considerably? A Global or a local approach?

Contributions

We present a statistical study of the suitability of global and local approaches for addressing automated face gender classification under realistic conditions. Main characteristics of this study:

- Cross-database experiments involving 3 different databases.
- Classifiers: 1-NN, PCA+LDA, SVM.
- Features: Grey levels and PCA.
- Statistical analysis of the results using several statistical tests.

Methodology

Given a test image, the next 3 steps are followed to classify it as male or female.

1. **Image Preprocessing**
   - detects the face in the image,
   - equalizes it,
   - and resizes it.
2. **Feature Extraction**
   - Grey levels or PCA feature vectors.
   - Global: Holistic descriptions.
   - Local: Description per patches.
3. **Classification**
   - Global: Classifiers work as usual.
   - Local: Combination of decisions based on patches.

Conclusions

In realistic scenarios:

- Global and local approaches achieve statistically equal accuracies.

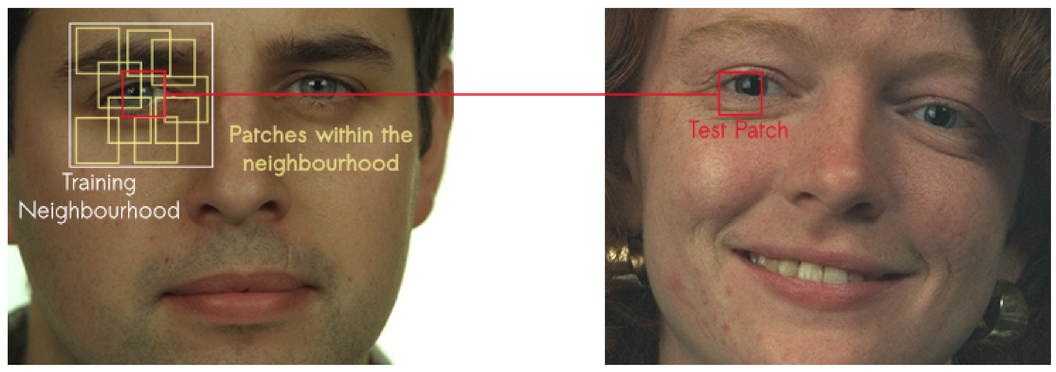
When the training and test images share the same characteristics and acquisition conditions:

- Global approaches perform better than local ones.
- A global SVM using grey levels is more likely to obtain the highest classification accuracies.

init
institute of new imaging technologies

Local vs Global Approaches

- **Global Approach**: Faces are described as a whole. Then, classified as usual.
- **Local Approach**: Faces are described per patches. Then, for each test patch its gender is estimated by comparing it with the patches from a neighbourhood in the training set. Finally, the gender of the face is predicted by majority voting of the local decisions.



The patches are overlapping with 1 pixel shift from one patch to its neighbours.

Classification Results

Gender classification accuracies (%) obtained in all experiments:

Training Data Set	Test Data Set	Global NN Grey Levels	Global NN PCA	Global PCA+LDA	Global SVM Grey Levels	Global SVM PCA	Local NN Grey Levels	Local NN PCA	Local PCA+LDA
FERET	FERET	85.31	85.57	91.86	93.66	92.83	92.35	91.29	85.07
	PAL	66.03	64.98	71.25	66.72	62.55	66.03	62.19	60.80
	AR Neutral	79.17	82.31	77.69	81.54	84.62	86.15	86.92	83.08
PAL	FERET	66.53	65.56	75.22	72.99	70.66	63.16	62.07	77.11
	PAL	77.42	77.35	82.72	85.23	85.61	83.73	83.52	73.69
	AR Neutral	81.25	82.31	89.23	92.31	91.54	90.00	90.00	87.69
AR Neutral	FERET	76.02	76.86	80.09	80.83	77.21	78.90	78.90	78.20
	PAL	73.35	72.30	71.43	75.09	70.38	74.39	73.17	65.51
	AR Neutral	83.99	82.46	87.54	90.42	98.15	88.92	89.08	86.31

Statistical Analysis of the Results

- Statistical analysis of the accuracies of **all experiments**:

Iman-Davenport's Statistic

$F_F = 12.18$
$F(7,35)_{0.95} = 2.29$

Differences were found

Holm's Method

1NN-pca-G	0.007143
1NN-grey-G	0.008333
PCALDA-L	0.01
1NN-pca-L	0.0125
PCALDA-G	0.016667
SVM-pca-G	0.025
1NN-grey-L	0.05

Rejected

SVM-grey-G

Wilcoxon's Test

	1	2	3	4	5	6	7	8
1NN-grey-G (1)	-		○	○	○	○	○	
1NN-pca-G (2)		-	○	○	○	○	○	
PCALDA-G (3)	●	●	-					
SVM-grey-G (4)	●	●		-		●	●	●
SVM-pca-G (5)	●				-			●
1NN-grey-L (6)	●	●				-		
1NN-pca-L (7)							-	
PCALDA-L (8)				○				-

Global methods seem to be more suitable than local ones.

- Statistical analysis of the accuracies of **only cross-database experiments**:

Iman-Davenport's Statistic

$F_F = 1.53$
$F(7,35)_{0.95} = 2.29$

No differences were found

Holm's Method

1NN-pca-G	0.007143
1NN-grey-G	0.008333
PCALDA-L	0.01
SVM-pca-G	0.0125
1NN-pca-L	0.016667
PCALDA-G	0.025
1NN-grey-L	0.05

Rejected

1NN-grey-L

Wilcoxon's Test

	1	2	3	4	5	6	7	8
1NN-grey-G (1)	-		○					
1NN-pca-G (2)		-	○					
PCALDA-G (3)			-					
SVM-grey-G (4)	●			-				
SVM-pca-G (5)					-			
1NN-grey-L (6)						-		
1NN-pca-L (7)							-	
PCALDA-L (8)								-

No major differences were found among the performances of the classification models.

Key to interpret these results:

- Iman-Davenport's Statistic (F_F) is higher than the corresponding value of the F-distribution when statistical differences are found.
- Holm's Method: The classification models above the double line performed significantly worse than the most significant model (marked in bold at the bottom) with a 95% significance level.
- Wilcoxon's Test: The symbol "●" indicates that the classification model in the row significantly outperforms the model in the column, and viceversa for the symbol "○" (above the main diagonal with a 90% confidence level, and below it with a 95%).