DeepGender: Occlusion and Low Resolution Robust Facial Gender Classification via Progressively Trained Convolutional Neural Networks with Attention

Felix Juefei Xu*, Eshan Verma*, Parag Goel, Anisha Cherodian, and Marios Savvides

*Authors contribute equally.
Motivation

Attention based model [1]:

**Encoder**: CNN as feature extractor.

**Decoder**: RNN w/ LSTM, learns attention mechanism.

Visualization: the network can automatically fix its gaze on the salient objects (regions) in the image while generating the image caption word by word.

**Q: Can we control/enforce the attention shift in CNN?**

---

Motivation

From previous work, we know that the periocular region provides the most important **cues** for determining gender information.

The periocular region is also the **most salient** region on human faces.
Motivation

Q: How can we let the CNN shift its attention towards the periocular region, where gender classification has been proven to be the most effective?

The answer comes from our day-to-day experience with photography.

The sharp foreground object attracts the most attention in the saliency heat map.

As opposed to the blurred / out-of-focus background content.
Motivation

We should be able to answer these following questions first before designing the progressive training paradigm.

Q: How can we let the CNN shift its attention towards the periocular region, where gender classification has been proven to be the most effective? (previous slide)

Q: Why not just use the periocular region crop?

Q: Why blurring instead of blackening out?

Q: Why not let CNN directly learn the blurring step?
Motivation

Q: Why not just use the periocular region crop?

Although periocular region is the best for gender classification, we still want to resort to other facial parts (beard/moustache) for providing valuable gender cues. Especially true when periocular region is less ideal (sunglasses).

To strike a good balance between full face-only and periocular-only models, we carry out a progressive training paradigm for CNN that starts with the full face, and progressively zoom into the periocular region by leaving other facial regions blurred.

Hope the network is sufficiently generalized.
Motivation

Q: *Why blurring instead of blackening out?*

We just want to *steer the focus*, rather than completely eliminate the background. Blackening would create abrupt edges.

When blurred, *low frequency information* is still well preserved. One can still recognize the content of the image, e.g., dog, human face, objects, etc. from a blurred image.

Blurring outside the periocular region, and leaving the high frequency details at the periocular region will both help providing *global and structural context* of the image, as well as keeping the *minute details* intact at the region of interest.
Motivation

Q: Why not let CNN directly learn the blurring step?

CNN filters operate on the entire image, and blurring only part of the image is a pixel location dependent operation and thus is difficult to emulate in the CNN framework.
Enforcing Attention in the Training Images

We heuristically choose 7 blur levels, including the one with no blur at all.

Gaussian blur kernel with $\sigma = 7$.

Doing this is conceptually enforcing the network attention in the training images without the need of changing the network architecture.
Training starts with the first epoch group (Epoch Group 0, images with no blur), and the first CNN model $M_0$ is obtained and frozen after convergence. Then, we input the next epoch group for tuning the $M_0$ and in the end produce the second model $M_1$. Sequentially obtain models: $M_1$ to $M_k$.

AlexNet, 2-way softmax.

Each $M_j$ ( $j=0, \ldots, k$ ) is trained with 1000 epochs, with a batchsize of 128.
Implicit Low-Rank Regularization in CNN

We have shown that the low-pass filtering in Fourier analysis is closely related to the low-rank approximation in SVD.

In the context of this work, progressively training the CNN using blurred images serves as an implicit low-rank regularizer.

This phenomenon is loosely observed through the visualization of the trained filters, which will be further analyzed and studied in future work.
Database

Training set: sourced from 5 different datasets. (Table 2)

Dimension: 168x210

Testing set: Pinellas County Sheriff’s Office (PCSO) database, we use 400K out of 1.4M. To be added occlusion and low-res degradations.

Dimension: 168x210

Table 2: Datasets used for progressive CNN training.

<table>
<thead>
<tr>
<th>DB Name</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>JNET</td>
<td>1900</td>
<td>1371</td>
</tr>
<tr>
<td>mugshotDB</td>
<td>1772</td>
<td>805</td>
</tr>
<tr>
<td>Pinellas Subset</td>
<td>13215</td>
<td>3394</td>
</tr>
<tr>
<td>pdx2</td>
<td>46346</td>
<td>12402</td>
</tr>
<tr>
<td>olympic2012</td>
<td>4164</td>
<td>3634</td>
</tr>
<tr>
<td>Total</td>
<td>67397</td>
<td>21606</td>
</tr>
<tr>
<td></td>
<td></td>
<td>89003</td>
</tr>
</tbody>
</table>
Pre-processing on 400K PCSO Testing Images

Row 1: random missing pixel occlusions

Row 2: random additive Gaussian noise occlusions

Row 3: random contiguous occlusions

Percentage of degradation for Row 1-3: 10%, 25%, 35%, 50%, 65%, 75%.

Row 4: various zooming factors (2x, 4x, 8x, 16x) for low-resolution degradations
Experiment 1: Occlusion Robustness

Experiments on the 400K PCSO mugshot database (artificial occlusions)

(1) Random missing pixels occlusions
(2) Random additive Gaussian noise occlusions
(3) Random contiguous occlusions
Baseline on Clean Images

This is the gender classification accuracies on the 400K PCSO database.

Images are clean, without artificially added degradations.

As expected, if the testing images are clean, it is preferable to use $M_F$ rather than $M_P$.

$M_F$ corresponds to the model trained on full face (equivalent to $M_0$), and $M_P$ is one trained using only periocular region (last Epoch Group only). $M_1$-$M_6$ are the incremental models trained.
Random Missing Pixels Occlusions

$M_5$ performs the best with $M_6$ showing a dip, suggesting a tighter periocular region is not well-suited for such application.

Notice a flip in performance of $M_F$ and $M_P$ going from the 10% to 25% with the periocular model generalizing better for higher corruptions. The trend of improving performance between progressively trained models is maintained.

Table 3: Overall classification accuracy on the PCSO (400K). Images are corrupted with random missing pixels of various percentages.

<table>
<thead>
<tr>
<th>Corrup.</th>
<th>0%</th>
<th>10%</th>
<th>25%</th>
<th>35%</th>
<th>50%</th>
<th>65%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_F$</td>
<td>97.66</td>
<td>97.06</td>
<td>93.61</td>
<td>89.15</td>
<td>82.39</td>
<td>79.46</td>
<td>77.4</td>
</tr>
<tr>
<td>$M_1$</td>
<td>97.63</td>
<td>96.93</td>
<td>92.68</td>
<td>87.99</td>
<td>81.57</td>
<td>78.97</td>
<td>77.2</td>
</tr>
<tr>
<td>$M_2$</td>
<td>97.46</td>
<td>96.83</td>
<td>93.19</td>
<td>89.17</td>
<td>83.03</td>
<td>80.06</td>
<td>77.68</td>
</tr>
<tr>
<td>$M_3$</td>
<td>97.4</td>
<td>96.98</td>
<td>94.06</td>
<td>90.65</td>
<td>84.79</td>
<td>81.59</td>
<td>78.56</td>
</tr>
<tr>
<td>$M_4$</td>
<td>97.95</td>
<td>97.63</td>
<td>95.63</td>
<td>93.1</td>
<td>87.96</td>
<td>84.41</td>
<td>80.22</td>
</tr>
<tr>
<td>$M_5$</td>
<td>97.52</td>
<td>97.26</td>
<td>95.8</td>
<td>94.07</td>
<td>90.4</td>
<td>87.39</td>
<td>83.04</td>
</tr>
<tr>
<td>$M_6$</td>
<td>97.6</td>
<td>97.29</td>
<td>95.5</td>
<td>93.27</td>
<td>88.8</td>
<td>85.57</td>
<td>81.42</td>
</tr>
</tbody>
</table>

Table 3: Overall classification accuracy on the PCSO (400K). Images are corrupted with random missing pixels of various percentages.
$M_4 - M_6$ perform best for medium noise. For high noise, $M_5$ is the most robust.

Just as before, as the noise increases, the trend undertaken by the performance of $M_F$ & $M_P$ and $M_5$ & $M_6$ is maintained and so is the performance trend of the progressively trained models.

Table 4: Overall classification accuracy on the PCSO (400K). Images are corrupted with additive Gaussian random noise of various percentages.

<table>
<thead>
<tr>
<th>Corrup.</th>
<th>0%</th>
<th>10%</th>
<th>25%</th>
<th>35%</th>
<th>50%</th>
<th>65%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_F$</td>
<td>97.66</td>
<td>97</td>
<td>94.03</td>
<td>91.19</td>
<td>86.47</td>
<td>83.43</td>
<td>79.94</td>
</tr>
<tr>
<td>$M_1$</td>
<td>97.63</td>
<td>96.93</td>
<td>94</td>
<td>91.26</td>
<td>87</td>
<td>84.27</td>
<td>81.15</td>
</tr>
<tr>
<td>$M_2$</td>
<td>97.46</td>
<td>96.87</td>
<td>94.43</td>
<td>92.19</td>
<td>88.75</td>
<td>86.44</td>
<td>83.33</td>
</tr>
<tr>
<td>$M_3$</td>
<td>97.4</td>
<td>97</td>
<td>95.18</td>
<td>93.27</td>
<td>89.93</td>
<td>87.55</td>
<td>84.16</td>
</tr>
<tr>
<td>$M_4$</td>
<td>97.95</td>
<td>97.67</td>
<td>96.45</td>
<td>95.11</td>
<td>92.43</td>
<td>90.28</td>
<td>87.06</td>
</tr>
<tr>
<td>$M_5$</td>
<td>97.52</td>
<td>97.29</td>
<td>96.25</td>
<td>95.21</td>
<td>93.21</td>
<td>91.65</td>
<td>89.12</td>
</tr>
<tr>
<td>$M_6$</td>
<td>97.6</td>
<td>97.32</td>
<td>96.04</td>
<td>94.77</td>
<td>92.46</td>
<td>90.8</td>
<td>88.08</td>
</tr>
<tr>
<td>$M_P$</td>
<td>95.75</td>
<td>95.59</td>
<td>94.85</td>
<td>94</td>
<td>92.43</td>
<td>91.15</td>
<td>88.74</td>
</tr>
</tbody>
</table>
(3) Random Contiguous Occlusions

The most realistic occlusions are the first few cases, others are extreme cases.

For the former cases, $M_1 - M_3$ are able to predict the classes with the highest accuracy.

Our scheme of focused saliency helps generalizing over occlusions.

**Table 5:** Overall classification accuracy on the PCSO (400K). Images are corrupted with random contiguous occlusions of various percentages.
Experiment 2: Low Resolution Robustness

Experiments are on the 400K PCSO mugshot database.

For cases 2x, 4x, and 8x, the trend between $M_1$-$M_6$ and their performance with respect to $M_F$ is maintained.

For 16x case, progressive models $M_1$-$M_6$ are still better than full face model $M_F$.

Table 7: Overall classification accuracy on the PCSO (400K). Images are down-sampled to a lower resolution with various zooming factors.

<table>
<thead>
<tr>
<th>Zooming Factor</th>
<th>1x</th>
<th>2x</th>
<th>4x</th>
<th>8x</th>
<th>16x</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_F$</td>
<td>97.66</td>
<td>97.55</td>
<td>96.99</td>
<td>94.19</td>
<td>87.45</td>
</tr>
<tr>
<td>$M_1$</td>
<td>97.63</td>
<td>97.48</td>
<td>96.91</td>
<td>94.76</td>
<td>87.41</td>
</tr>
<tr>
<td>$M_2$</td>
<td>97.46</td>
<td>97.31</td>
<td>96.73</td>
<td>94.77</td>
<td>88.82</td>
</tr>
<tr>
<td>$M_3$</td>
<td>97.4</td>
<td>97.2</td>
<td>96.37</td>
<td>93.5</td>
<td>87.57</td>
</tr>
<tr>
<td>$M_4$</td>
<td>97.95</td>
<td>97.89</td>
<td>97.56</td>
<td>95.67</td>
<td>90.17</td>
</tr>
<tr>
<td>$M_5$</td>
<td>97.52</td>
<td>97.4</td>
<td>96.79</td>
<td>95.26</td>
<td>89.66</td>
</tr>
<tr>
<td>$M_6$</td>
<td>97.6</td>
<td>97.51</td>
<td>97.05</td>
<td>95.42</td>
<td>90.79</td>
</tr>
<tr>
<td>$M_F$</td>
<td>95.75</td>
<td>95.65</td>
<td>95.27</td>
<td>94.12</td>
<td>91.59</td>
</tr>
</tbody>
</table>
Conclusion and Discussion

The intuition:

(1) To have the network focus on the periocular region of the face for gender classification.

(2) To preserve contextual information of facial contours to generalize better over occlusions.

Our hypothesis is indeed true and that for a given occlusion set, it is possible to have high accuracy from a model that encompasses both of above stated properties.

We did not train on any occluded data, or optimize for a particular type of occlusions, our models can generalize well.
Future Work

We have observed significant testing time savings from M0 to M6. We have thus visualize the learned filters. It seems that after progressive training, the filters are smoother, and we will study the connection between the two in the future.
Thank you!

Questions?

Check out the poster.