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
CyLab Biometrics Center, Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{felixu,everma}@cmu.edu, {paraggoe,acherodi}@andrew.cmu.edu, msavvid@ri.cmu.edu
*Authors contribute equally.

Abstract

*In this work, we have undertaken the task of occlusion and low-resolution robust facial gender classification. Inspired by the trainable attention model via deep architecture, and the fact that the periocular region is proven to be the most salient region for gender classification purposes, we are able to design a progressive convolutional neural network training paradigm to enforce the attention shift during the learning process. The hope is to enable the network to attend to particular high-profile regions (*e.g.* the periocular region) without the need to change the network architecture itself. The network benefits from this attention shift and becomes more robust towards occlusions and low-resolution degradations. With the progressively trained CNN models, we have achieved better gender classification results on the large-scale PCSO mugshot database with 400K images under occlusion and low-resolution settings, compared to the one undergone traditional training. In addition, our progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations, as well as low resolution.*

1. Introduction

Facial gender classification has always been one of the most studied soft-biometric topics. Over the past decade, gender classification on constrained faces has almost been perfected. However, challenges still remain on less constrained faces such as faces with occlusions, of low resolution, and off-angle poses. Traditional methods such as the support vector machines (SVMs) and its kernel extension can work pretty well on this classic two-class problem as listed in Table 8. In this work, we approach this problem from a very different angle. We are inspired by the booming deep convolutional neural network (CNN) and the attention model to achieve occlusion and low resolution robust facial gender classification via progressively training the CNN with attention.

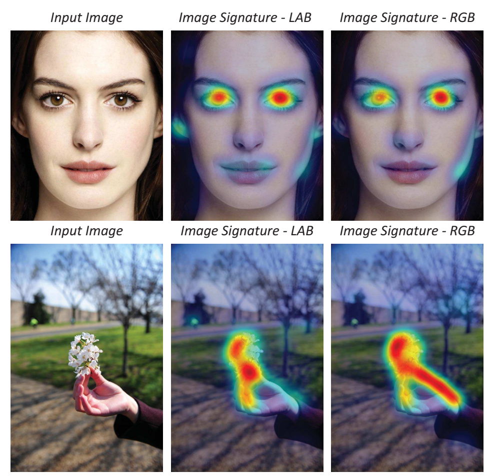


Figure 1: (Top) Periocular region on human faces exhibits the highest saliency. (Bottom) Foreground object in focus exhibits the highest saliency. Background is blurred with less high-frequency details preserved.

Motivation: Xu *et al*. [58] proposed an attention based model that automatically learns to describe the content of images which has been inspired by recent work in machine translation [2] and object detection [1, 48]. In their work, two attention-based image caption generators were introduced under a common framework: (1) a 'soft' deterministic attention mechanism which can be trained by standard back-propagation method and (2) a 'hard' stochastic attention mechanism which can be trained by maximizing an approximate variational lower bound. The encoder of the models uses a convolutional neural network as a feature extractor, and the decoder is comprised of a recurrent neural network (RNN) with long short-term memory (LSTM) architecture where the attention mechanism is learned. The authors can then visualize that the network can automatically fix its gaze on the salient objects (regions) in the image while generating the image caption word by word.

For facial gender classification, we know from previous work [13, 47] that the periocular region provides the most important cues for determining the gender information. The periocular region is also the most salient region on human faces, such as shown in the top part of Figure 1, using a general purpose saliency detection algorithm [11]. Similar results can also be obtained using other saliency detection algorithms such as [14] and [12]. We can observe from the saliency heat map that the periocular region does fire the most strongly compared to the remainder of the face.

Now we come to think about the following question:

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. If you are using a DSLR camera with a big aperture lens, and fixing the focal point onto an object in the foreground, all background beyond the object in focus will become out of focus and blurred. This is illustrated in the bottom part of Figure 1 and as can be seen, the sharp foreground object (cherry blossom in hand) attracts the most attention in the saliency heat map.

Thus, we can control the attention region by specifying where the image is blurred or remains sharp. In the context of gender classification, we know that we can benefit from fixing the attention onto the periocular region. Therefore, we are 'forcing' what part of the image the network weighs the most, by progressively training the CNN using images with increasing blur levels, zooming into the periocular region, as shown in Table 1. Since we still want to use a full face model, we hope that by employing the mentioned strategy, the learned deep model can be at least on par with other full face deep models, while harnessing gender cues in the periocular region.

Q: Why not just use the periocular region crop?

Although experimentally, periocular is the best facial region for gender classification, we still want to resort to other facial parts (beard/moustache) for providing valuable gender cues. This is specially true when the periocular region is less ideal. For example, some occlusion like sunglasses could be blocking the eye region, and we want our network to still be able to generalize well and perform robustly, even when the periocular region is corrupted.

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. In addition, we hope that the progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations.

Q: Why blurring instead of blackening out?

We just want to steer the focus, rather than completely eliminating the background, like the DSLR photo example shown in the bottom part of Figure 1. Blackening would create abrupt edges that confuse the filters during the training. 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, which will help the gender classification, and fine-grained categorization in general.

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

We know that 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. Therefore, we carry out the proposed progressive training paradigm to enforce where the network attention should be.

2. Related Work

We provide relevant background on facial gender classification and attention models.

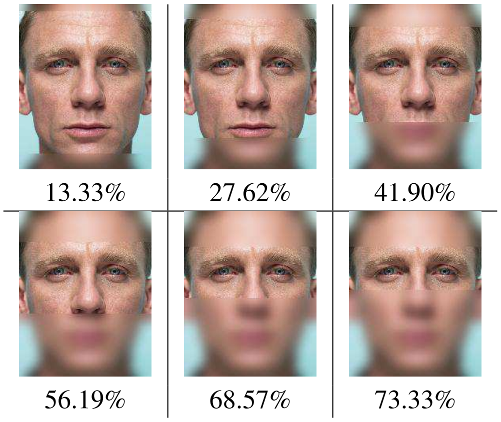
The periocular region is shown to be the best facial region for recognition purposes [37, 20, 23, 33, 22, 30, 27, 23, 51, 21, 26, 19, 18, 28, 50], especially for gender classification tasks [13, 47]. A few recent work also applies CNN for gender classification [41] and [3]. More related work on gender classification is consolidated in Table 8.

Attention models such as the one used for image captioning [58] have gained much popularity only very recently. Rather than compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. It also helps gaining insight and interpreting the results by visualizing where the model attends to for certain tasks. This mechanism can be viewed as a learnable saliency detector that can be tailored to various tasks, as opposed to the traditional ones such as [11, 14, 12, 6].

It is worth mentioning the key difference between the soft attention and the hard attention. The soft attention is very easy to implement. It produces distribution over input locations, re-weights features and feeds them as input. It can attend to arbitrary input locations using spatial transformer networks [16]. On the other hand, the hard attention can only attend to a single input location, and the optimization cannot utilize gradient descent. The common practice is to use reinforcement learning.

Other applications involving attention models may include machine translation which applies attention over in-

Table 1: Blurred images with increasing levels of blur.



put [45]; speech recognition which applies attention over input sounds [4, 7]; video captioning with attention over input frames [60]; image, question to answer with attention over image itself [57, 62]; and many more [55, 56].

3. Proposed Method

In this section we detail our proposed method on progressively training the CNN with attention. The entire training procedure involves $(k+1)$ epoch groups from epoch group 0 to k, where each epoch group corresponds to one particular blur level.

3.1. Enforcing Attention in the Training Images

In our experiment, we heuristically choose 7 blur levels, including the one with no blur at all. The example images with increasing blur levels are illustrated in Table 1. We use a Gaussian blur kernel with $\sigma = 7$ to blur the corresponding image regions. Doing this is conceptually enforcing the network attention in the training images without the need of changing the network architecture.

3.2. Progressive CNN Training with Attention

We employ the AlexNet [38] architecture for our progressive CNN training. The AlexNet has 60 million parameters and 650,000 neuron, consisting of 5 convolution layers and 3 fully connected layers with a final 1000-way softmax. To reduce overfitting in the fully-connected layers, AlexNet employs “dropout” and data-augmentation, both of which are preserved in our training. The main difference is that we only need a 2-way softmax due to the nature of gender classification problems.

As illustrated in Figure 2, the progressive CNN training begins with the first epoch group (Epoch Group 0, images with no blur), and the first CNN model $\mathcal{M}_0$ is obtained and frozen after convergence. Then, we input the next epoch group for tuning the $\mathcal{M}_0$ and in the end produce the second model $\mathcal{M}_1$, with attention enforced through training images. The procedure is carried out sequentially until the final model $\mathcal{M}_k$ is obtained. Each $\mathcal{M}_j (j = 0, \ldots, k)$ is trained with 1000 epochs and with a batchsize of 128. At the end of the training for step j, the model corresponding to best validation accuracy is taken ahead to the next iteration $(j+1)$.

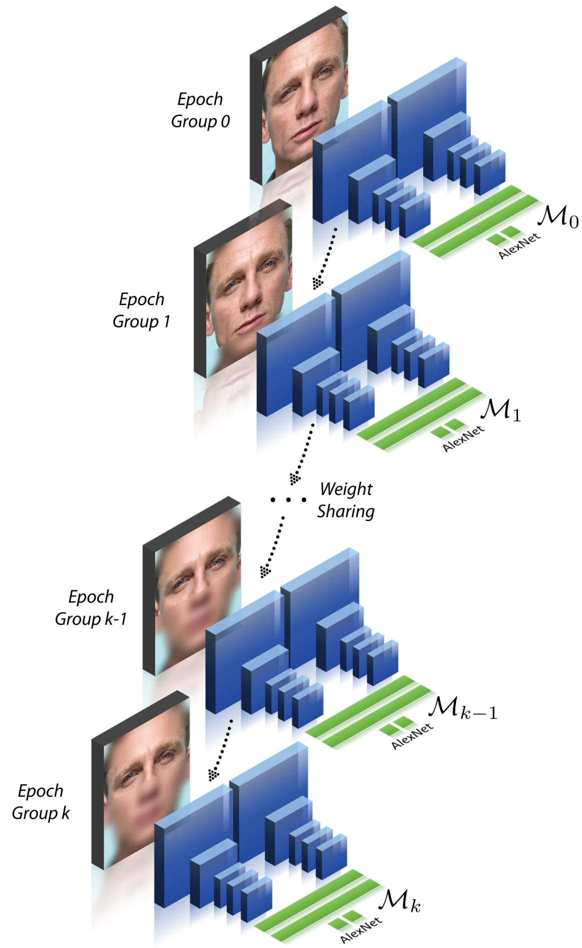


Figure 2: Progressive CNN training with attention.

3.3. Implicit Low-Rank Regularization in CNN

Blurring the training images in our paradigm may have more implications. Here we want to show the similarities between low-pass Fourier analysis and low-rank approximation in SVD. Through the analysis, we hope to make connections to the low-rank regularization procedure in the CNN. We have learned from a recent work [53] that enforcing a low-rank regularization and removing the redundancy in the convolution kernels is important and can help improve both the classification accuracy and the computation speed. Fourier analysis involves expansion of the original data x_{ij} (taken from the data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$) in an orthogonal basis, which is the inverse Fourier transform:

$$x_{ij} = \frac{1}{m}\sum_{k=0}^{m-1} c_k e^{\mathbf{i}2\pi jk/m} \tag{1}$$

The connection with SVD can be explicitly illustrated by normalizing the vector $\{e^{\mathbf{i}2\pi jk/m}\}$ and by naming it $\mathbf{v}'_k$:

$$x_{ij} = \sum_{k=0}^{m-1} b_{ik}v'_{jk} = \sum_{k=0}^{m-1} u'_{ik}s'_k v'_{jk} \tag{2}$$

which generates the matrix equation $\mathbf{X} = \mathbf{U}'\mathbf{\Sigma}'\mathbf{V}'^{\top}$. However, unlike the SVD, even though the $\{\mathbf{v}'_k\}$ are an orthonormal basis, the $\{\mathbf{u}'_k\}$ are not in general orthogonal. Nevertheless this demonstrates how the SVD is similar to a Fourier transform. Next, we will show that the **low-pass filtering in Fourier analysis** is closely related to the **low-rank approximation in SVD**.

Suppose we have N image data samples in original two-dimensional form $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and each has dimension d. Let matrix $\hat{\mathbf{X}}$ contain all the data samples undergone 2D Fourier transform $\mathcal{F}(\cdot)$, in the vectorized form:

$$\hat{\mathbf{X}} = \begin{bmatrix} | & | & & | \\ \mathrm{vec}(\mathcal{F}(\mathbf{x}_1)) & \mathrm{vec}(\mathcal{F}(\mathbf{x}_2)) & \ldots & \mathrm{vec}(\mathcal{F}(\mathbf{x}_N)) \\ | & | & & | \end{bmatrix}_{d\times N}$$

Matrix $\hat{\mathbf{X}}$ can be decomposed using SVD: $\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\top}$. Without loss of generality, let us assume that $N = d$ for brevity. Let $\mathbf{g}$ and $\hat{\mathbf{g}}$ be the Gaussian filter in spatial domain and frequency domain respectively, namely $\hat{\mathbf{g}} = \mathcal{F}(\mathbf{g})$. Let $\hat{\mathbf{G}}$ be a diagonal matrix with $\hat{\mathbf{g}}$ on its diagonal. The convolution operation becomes dot product in frequency domain, so the blurring operation becomes:

$$\hat{\mathbf{X}}_{\text{blur}} = \hat{\mathbf{G}}\cdot\hat{\mathbf{X}} = \hat{\mathbf{G}}\cdot\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\top} \tag{3}$$

where $\hat{\mathbf{\Sigma}} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$ contains the singular values of $\hat{\mathbf{X}}_{\text{blur}}$, already sorted in descending order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$. Suppose we can find a permutation matrix $\mathbf{P}$ such that when applied on the diagonal matrix $\hat{\mathbf{G}}$, the diagonal elements is sorted in descending order according to the magnitude: $\hat{\mathbf{G}}' = \mathbf{P}\hat{\mathbf{G}} = \mathrm{diag}(\hat{g}'_1, \hat{g}'_2, \ldots, \hat{g}'_d)$. Now, let us apply the same permutation operation on $\hat{\mathbf{X}}_{\text{blur}}$, we can thus have the following relationship:

$$\mathbf{P}\cdot\hat{\mathbf{X}}_{\text{blur}} = \mathbf{P}\cdot\hat{\mathbf{G}}\cdot\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\top} \tag{4}$$

$$\hat{\mathbf{X}}'_{\text{blur}} = \hat{\mathbf{G}}'\cdot\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{\top} = \hat{\mathbf{U}}\cdot(\hat{\mathbf{G}}'\hat{\mathbf{\Sigma}})\cdot\hat{\mathbf{V}}^{\top} \tag{5}$$

$$= \hat{\mathbf{U}}\cdot\mathrm{diag}(\hat{g}'_1\sigma_1, \hat{g}'_2\sigma_2, \ldots, \hat{g}'_d\sigma_d)\cdot\hat{\mathbf{V}}^{\top} \tag{6}$$

Due to the fact that Gaussian distributing is not a heavy-tailed distribution, the already smaller singular values will be brought down to 0 by the Gaussian weights. Therefore, $\hat{\mathbf{X}}_{\text{blur}}$ actually becomes low-rank after Gaussian low-pass filtering. To this end, we can say that low-pass filtering in Fourier analysis is equivalent to the low-rank approximation in SVD up to a permutation.

This phenomenon is loosely observed through the visualization of the trained filters, as shown in Figure 10, which will be further analyzed and studied in future work.

Table 2: Datasets used for progressive CNN training.

DB Name	Males	Females
JNET	1900	1371
mugshotDB	1772	805
Pinellas Subset	13215	3394
pdx2	46346	12402
olympic2012	4164	3634
Total	67397	21606
	89003	

4. Experiments

In this section we detail the training and testing protocols employed and various occlusions and low resolutions modeled in the testing set. Accompanying figures and tables for each sub-section encompass the results and observations and are elaborated in each section.[1]

4.1. Database and Pre-processing

Training set: We source images from 5 different datasets, each containing samples of both classes. The datasets are JNET, olympic2012, mugshotDB, pdx2 and Pinellas. All the datasets, except Pinellas are evenly separated into males and females of different ethnicity. The images are centred. By which, we mean that we have landmarked certain points on the face, which are then anchored to fixed points in the resulting training image. For example, the eyes are anchored at the same coordinates in every image. All of our input images have the same dimension 168×210. The details of the training datasets are listed in Table 2. The images are partitioned into training and validation and the progressive blur is applied to each image as explained in the previous section. Hence, for a given model iteration, the training set consists of ∼90k images.

Testing set: The testing set was built primarily from the following two datasets: (1) The AR Face database [46] is one of the most widely used face databases with occlusions. It contains 3,288 color images from 135 subjects (76 male subjects + 59 female subjects). Typical occlusions include sunglasses and scarves. The database also captures expression variations and lighting changes. (2) Pinellas County

[1] **A note on legend**: (1) Symbols $\mathcal{M}$ correspond to each model trained, with $\mathcal{M}_F$ corresponding to the model trained on full face (equivalent to $\mathcal{M}_0$), $\mathcal{M}_P$ to one with just periocular images and $\mathcal{M}_k$, $k \subseteq (1, \ldots, 6)$ to the incremental models trained. (2) The tabular results show model performance on the original images in column 1 and corrupted images in other columns.

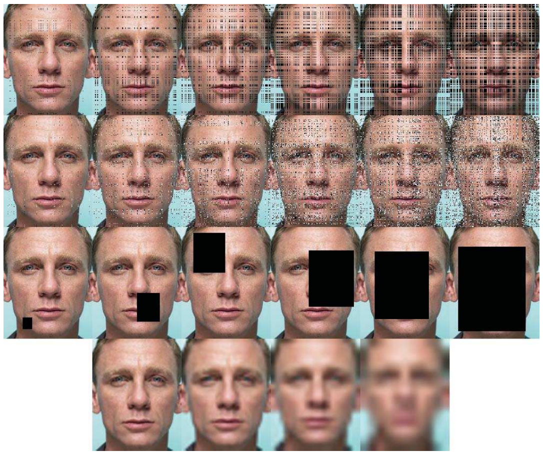


Figure 3: Various degradations applied on the testing images. Row 1: random missing pixel occlusions; Row 2: random additive Gaussian noise occlusions; Row 3: random contiguous occlusions. Percentage of degradation for Row 1-3: 10%, 25%, 35%, 50%, 65%, 75%. Row 4: various zooming factors (2x, 4x, 8x, 16x) for low-resolution degradations.

Sherrif's Office (PCSO) mugshot database is a large-scale database of over 1.4 million images. We took a subset of around 400K images from this dataset. These images are not seen during training.

The testing images are centered and cropped in the same way as the training images, though other pre-processing like the progressive blur are not applied. Instead, to model real world occlusions we have conducted the following experiments to be discussed in Section 4.2.

4.2. Experiment I: Occlusion Robustness

In Experiment I, we carry out occlusion robust gender classification on both the AR Face database and the PCSO mugshot database. We manually add artificial occlusions to test the efficacy of our method on the PCSO database and test on the various images sets in the AR Face dataset.

Experiments on the PCSO mugshot database:

To begin with, the performance of various models on the clean PCSO data is shown in Figure 4. As expected, if the testing images are clean, it should be preferable to use $\mathcal{M}_F$, rather than $\mathcal{M}_P$. We can see that the progressively trained models $\mathcal{M}_1 - \mathcal{M}_6$ are on par with $\mathcal{M}_F$.

We corrupt the testing images (400K) with three types of facial occlusions. These are visualized in Figure 3 with each row corresponding to some modeled occlusions.

(1) Random missing pixels occlusions: Varying factors of the image pixels (10%, 25%, 35%, 50%, 65%, 75%) were dropped to model lost information and grainy images[2]. This is corresponding to the first row in Figure 3. From Table 3 and Figure 5, $\mathcal{M}_5$ performs the best with $\mathcal{M}_6$ showing a dip in accuracy suggesting a tighter periocular region is not well suited for such applications, *i.e.*, a limit on the periocular region needs to be maintained in the blur-set. There is a flip in performance of the models $\mathcal{M}_P$ and $\mathcal{M}_F$ going from the original to 25% with the periocular model generalizing better for higher corruptions. As the percentage of missing pixels increases, the performance gap between $\mathcal{M}_P$ and $\mathcal{M}_F$ increases. As hypothesized, the trend of improving performance between progressively trained models is maintained across factors indicating a better learned model towards noise.

(2) Random additive Gaussian noise occlusions: Gaussian white noise ($\sigma = 6$) was added to image pixels in varying factors (10%, 25%, 35%, 50%, 65%, 75%). This is corresponding to the second row in Figure 3 and is done

[2]This can also model the dead pixel/shot noise of a sensor and these results can be used to accelerate in-line gender detection by using partially demosaiced images.

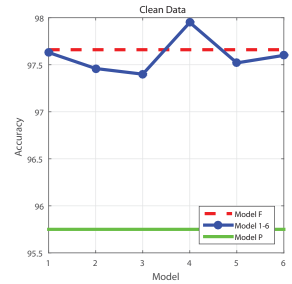


Figure 4: Overall classification accuracy on the PCSO (400K). Images are not corrupted.

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

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97.06	93.61	89.15	82.39	79.46	77.4
$\mathcal{M}_1$	97.63	96.93	92.68	87.99	81.57	78.97	77.2
$\mathcal{M}_2$	97.46	96.83	93.19	89.17	83.03	80.06	77.68
$\mathcal{M}_3$	97.4	96.98	94.06	90.65	84.79	81.59	78.56
$\mathcal{M}_4$	97.95	97.63	95.63	93.1	87.96	84.41	80.22
$\mathcal{M}_5$	97.52	97.26	95.8	94.07	90.4	87.39	83.04
$\mathcal{M}_6$	97.6	97.29	95.5	93.27	88.8	85.57	81.42
$\mathcal{M}_P$	95.75	95.45	93.84	92.02	88.87	86.59	83.18

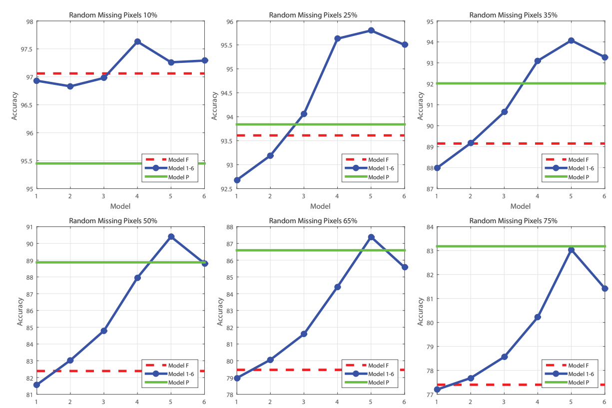


Figure 5: Overall classification accuracy on the PCSO (400K). Images are corrupted with **random missing pixels** of various percentages.

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

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97	94.03	91.19	86.47	83.43	79.94
$\mathcal{M}_1$	97.63	96.93	94	91.26	87	84.27	81.15
$\mathcal{M}_2$	97.46	96.87	94.43	92.19	88.75	86.44	83.33
$\mathcal{M}_3$	97.4	97	95.18	93.27	89.93	87.55	84.16
$\mathcal{M}_4$	97.95	97.67	96.45	95.11	92.43	90.28	87.06
$\mathcal{M}_5$	97.52	97.29	96.25	95.21	93.21	91.65	89.12
$\mathcal{M}_6$	97.6	97.32	96.04	94.77	92.46	90.8	88.08
$\mathcal{M}_P$	95.75	95.59	94.85	94	92.43	91.15	88.74

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

Corrup.	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	96.69	93.93	88.63	76.54	73.75	64.82
$\mathcal{M}_1$	97.63	96.95	94.64	90.2	77.47	75.2	53.04
$\mathcal{M}_2$	97.46	96.76	94.56	90.04	75.99	70.83	56.25
$\mathcal{M}_3$	97.4	96.63	94.65	90.08	77.13	71.77	68.52
$\mathcal{M}_4$	97.95	96.82	92.7	86.64	75.25	70.37	61.63
$\mathcal{M}_5$	97.52	96.56	92.03	83.95	70.36	69.94	66.52
$\mathcal{M}_6$	97.6	96.61	93.08	86.34	71.91	71.4	69.5
$\mathcal{M}_P$	95.75	95	93.01	88.34	76.82	67.81	49.73

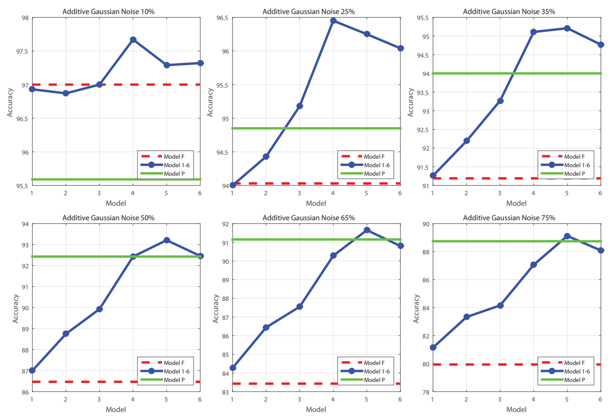


Figure 6: Overall classification accuracy on the PCSO (400K). Images are corrupted with **additive Gaussian random noise** of various percentages.

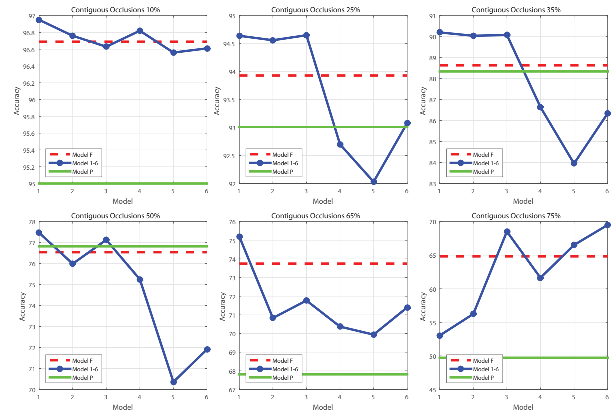


Figure 7: Overall classification accuracy on the PCSO (400K). Images are corrupted with **random contiguous occlusions** of various percentages.

to model high noise data and bad compression. From Table 4 and Figure 6, $\mathcal{M}_4 - \mathcal{M}_6$ perform best for medium noise. For high noise, $\mathcal{M}_5$ is the most robust. Just like before, as the noise increases, the trend undertaken by the performance of $\mathcal{M}_P$ & $\mathcal{M}_F$ and $\mathcal{M}_5$ & $\mathcal{M}_6$ is maintained and so is the performance trend of the progressively trained models.

(3) Random contiguous occlusions: To model big occlusions like sunglasses or other contiguous elements, continuous patches of pixels (10%, 25%, 35%, 50%, 65%, 75%) were dropped from the image as seen in the third row of Figure 3. The most realistic occlusion correspond to the first few patches, other patches are extreme cases. For the former cases, $\mathcal{M}_1 - \mathcal{M}_3$ are able to predict the classes with the highest accuracy. From Table 5 and Figure 7, for such large occlusions and missing data, more contextual information is needed for correct classification since $\mathcal{M}_1 - \mathcal{M}_3$ perform better than other models. However, since they perform better than $\mathcal{M}_F$, our scheme of focused saliency helps generalizing over occlusions.

Experiments on the AR Face database:

We partitioned the original set to smaller subsets to better understand our methodology's performance under different conditions. Set 1 consists of neutral expression, full-face subjects. Set 2 has full-face but varied expressions. Set 3 includes periocular occlusions such as sunglasses and Set 4 includes these and other occlusions like clothing *etc*. Set 5 is the entire dataset including illumination variations.

Referring to Table 6 and Figure 8, for Set 1, the full face model performs the best and this is expected as this model was trained on images very similar to this. Set 2 suggests that the models need more contextual information when expressions are introduced. Thus, $\mathcal{M}_4$ which has focus on periocular but has face information too performs best. For Set 3, we can see two things, one, $\mathcal{M}_P$ performs better than $\mathcal{M}_F$ indicative of its robustness to periocular occlusions. Two, $\mathcal{M}_5$ is the best as it combines periocular focus with contextual information gained from incremental training.

Set 4 performance brings out why periocular region is preferred for occluded faces. We ascertained that some texture and loss of face contour is throwing off the models $\mathcal{M}_1 - \mathcal{M}_6$. The performance of the models on Set 5 reiterates previously stated observations of the combined importance of contextual information about face contours and the importance of periocular region. This is the reason for the best accuracy reported by $\mathcal{M}_3$.

4.3. Experiment II: Low Resolution Robustness

Our scheme of training on Gaussian blurred images should generalize well to low resolution images. To test this hypothesis, we tested our models on images from the PCSO mugshot dataset by first down-sampling them by a factor and then blowing them back up (zooming factor for

Table 6: Gender classification accuracy on the AR Face database.

Sets	Set 1	Set 2	Set 3	Set 4	Set 5 (Full Set)
$\mathcal{M}_F$	98.44	93.23	89.06	83.04	81.65
$\mathcal{M}_1$	97.66	92.71	86.72	81.7	82.82
$\mathcal{M}_2$	97.66	92.71	90.62	82.14	85.1
$\mathcal{M}_3$	97.66	93.23	91.41	80.8	85.62
$\mathcal{M}_4$	98.44	95.31	92.97	77.23	84.61
$\mathcal{M}_5$	96.88	93.49	94.53	80.36	84.67
$\mathcal{M}_6$	96.09	92.71	92.97	79.02	83.9
$\mathcal{M}_P$	96.09	90.62	91.41	86.61	83.44

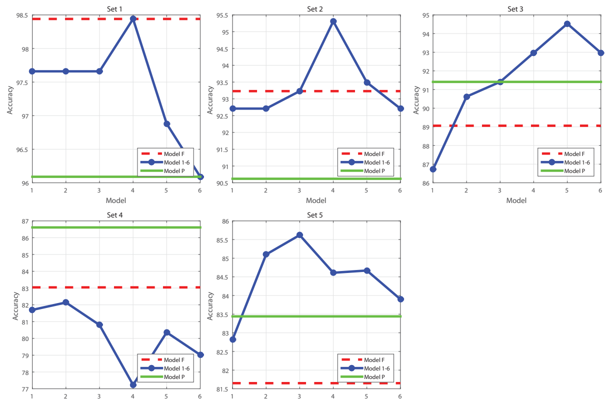


Figure 8: Gender classification accuracy on the AR Face database.

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

Zooming Factor	1x	2x	4x	8x	16x
$\mathcal{M}_F$	97.66	97.55	96.99	94.19	87.45
$\mathcal{M}_1$	97.63	97.48	96.91	94.76	87.41
$\mathcal{M}_2$	97.46	97.31	96.73	94.77	88.82
$\mathcal{M}_3$	97.4	97.2	96.37	93.5	87.57
$\mathcal{M}_4$	97.95	97.89	97.56	95.67	90.17
$\mathcal{M}_5$	97.52	97.4	96.79	95.26	89.66
$\mathcal{M}_6$	97.6	97.51	97.05	95.42	90.79
$\mathcal{M}_P$	95.75	95.65	95.27	94.12	91.59

example: 2x, 4x, 8x, 16x)[3]. This inculcates the loss of edge information and other higher order information and is captured in the last row of Figure 3. As seen in Table 7 and Figure 9 for cases, 2x, 4x, 8x, the trend between $\mathcal{M}_1 - \mathcal{M}_6$ and their performance with respect to $\mathcal{M}_F$ is maintained. As mentioned before, $\mathcal{M}_4$ performs well due to the balance between focus on periocular region and saving the contextual information of a face.

4.4. Summary and Discussion

We have proposed a methodology for building a gender recognition system which is robust to occlusions. It involves training a deep model incrementally over several batches of input data pre-processed with progressive blur. The intuition and intent is two-fold, one to have the network focus on periocular regions of the face for gender recognition. And two, to preserve contextual information of facial

[3]Effective pixel for 16x zooming factor is around 10x13, which is a quite challenging low-resolution setting.

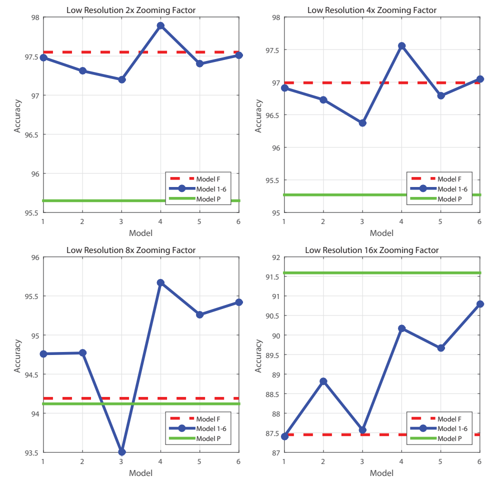


Figure 9: Overall classification accuracy on the PCSO (400K). Images are down-sampled to a **lower resolution** with various zooming factors.

contours to generalize better over occlusions.

Through various experiments we have observed that 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. Irrespective of the fact that we did not train on any occluded data, or optimize for a particular set of occlusions, our models are able to generalize well over synthetic data and real life facial occlusion images.

We have summarized the overall experiments and consolidated the results in Table 8. For PCSO large-scale experiments, we believe that 35% occlusion is the right amount of degradations, on which accuracies should be reported. Therefore we average the accuracy from our best model on three types of occlusions (missing pixel, additive Gaussian noise, and contiguous occlusions) which gives 93.12% in Table 8. For low-resolution experiments, we believe 8x zooming factor is the right amount of degradations, so we report the accuracy 95.67% in Table 8. Many other related work on gender classification are also listed for a quick comparison. This table is based on [13].

5. Conclusions and Future Work

In this work, we have undertaken the task of occlusion and low-resolution robust facial gender classification. Inspired by the trainable attention model via deep architecture, and the fact that the periocular region is proven to be the most salient region for gender classification purposes, we are able to design a progressive convolutional neural network training paradigm to enforce the attention shift during the learning process. The hope is to enable the network to attend to particular high-profile regions (*e.g.* the periocular region) without the need to change the network architecture itself. The network benefits from this attention shift and be-

Table 8: Summary of many related work on gender classification. The proposed method is shown in the top rows.

Authors	Methods	Dataset	Variation	Unique Subj.	Resolution	# of Subjects (Male/Female)	Accuracy
Proposed	Progressive CNN training w/ attention	Mugshots	Frontal-only, mugshot	Yes	168x210	89k total tr, 400k total te	97.95% te
			Occlusion	Yes		89k total tr, 400k total te	93.12% te
			Low-resolution	Yes		89k total tr, 400k total te	95.67% te
		AR Face	Expr x4, occl x2	No		89k total tr, 76/59 te	85.62% te
Hu *et al.* [13]	Region-based MR-8 filter bank w/ fusion of linear SVMs	Flickr	Li,exp,pos,bkgd,occl	Yes	128x170	10037/10037 tr, 3346/3346 te	90.1% te
		FERET	Frontal-only, studio	Yes		320/320 tr, 80/80 te	92.8% te
Chen & Lin [5]	Color & edge features w/ Adaboost+weak classifiers	Web img	Lighting, expression background	Yes	N/A	1948 total tr, 210/259 te	87.6% te
Wang *et al.* [8]	Gabor filters w/ polynomial-SVM	BioID	Lighting & expression	Yes	286x384	976/544 tr, 120 total te	92.5% te
Golomb *et al.* [9]	Raw pixel w/ neural network	SexNet	Frontal-only, studio	Yes	30x30	40/40 tr, 5/5 te	91.9% tr
Gutta *et al.* [10]	Raw pixel w/ mix of neural net RBF-SVM & decision tree	FERET	Frontal-only, studio	No	64x72	1906/1100 tr, 47/30 te	96.0% te
Jabid *et al.* [15]	Local directional patterns w/ SVM	FERET	Frontal-only, studio	No	100x100	1100/900 tr, unknown te	95.1% te
Lee *et al.* [39]	Region-based w/ linear regression fused w/ SVM	FERET	Frontal-only, studio	N/A	N/A	1158/615 tr, unknown te	98.8% te
		Web img	Unknown			1500/1500 tr, 1500/1500 te	88.1% te
Leng & Wang [40]	Gabor filters w/ fuzzy-SVM	FERET	Frontal-only, studio	Yes	256x384	160/140 total, 80% tr, 20% te	98.1% te
		CAS-PEAL	Studio	N/A	140x120	400/400 total, 80% tr, 20% te	93.0% te
		BUAA-IRIP	Frontal-only, studio	No	56x46	150/150 total, 80% tr, 20% te	89.0% te
Lin *et al.* [42]	Gabor filters w/ linear SVM	FERET	Frontal-only, studio	N/A	48x48	Unknown	92.8% te
Lu & Lin [43]	Gabor filters w/ Adaboost + linear SVM	FERET	Frontal-only, studio	N/A	48x48	150/150 tr, 518 total te	90.0% te
Lu *et al.* [44]	Region-based w/ RBF-SVM fused w/ majority vote	CAS-PEAL	Studio	Yes	90x72	320/320 tr, 80/80 te	92.6% te
Moghaddam & Yang [49]	Raw pixel w/ RBF-SVM	FERET	Frontal-only, studio	N/A	80x40	793/715 tr, 133/126 te	96.6% te
Yang *et al.* [59]	Texture normalization w/ RBF-SVM	Snapshots	Unknown	N/A	N/A	5600/3600 tr, unknown te	97.2% te
		FERET	Frontal-only, studio			1400/900 tr, 3529 total te	92.2% te

comes more robust towards occlusions and low-resolution degradations. With the progressively trained CNN models, we have achieved better gender classification results on the large-scale PCSO mugshot database with 400K images under occlusion and low-resolution settings, compared to the one undergone traditional training. In addition, our progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations, as well as low resolution.

Future work: We have carried out a set of large-scale testing experiments on the PCSO mugshot database with 400K images, shown in the experimental section. We have noticed that, under the same testing environment, the amount of time it takes to test on the entire 400K images various dramatically for different progressively trained models ($\mathcal{M}_0 - \mathcal{M}_6$). As shown in Figure 10, we can observe a trend of testing time decrease when testing using $\mathcal{M}_0$ all the way to $\mathcal{M}_6$, where the curves correspond to the additive Gaussian noise occlusion robust experiments. This same trend is observed across the board for all the large-scale experiments on PCSO. The time difference is stunning. For example, if we look at the green curve, $\mathcal{M}_0$ takes over 5000 seconds while $\mathcal{M}_6$ only around 500. One of the future directions is to study the cause of this phenomenon. One possible direction is to study the sparsity or the smoothness of the learned filters.

Shown in our visualization (Figure 10) of the 64 first-layer filters in AlexNet for models $\mathcal{M}_0$, $\mathcal{M}_3$, and $\mathcal{M}_6$, respectively, we can observe that the progressively trained filters seems to be smoother and this may be due to the implicit low-rank regularization phenomenon discussed in Section 3.3. Other future work may include studying how the ensemble of models can further improve the performance and how various multi-modal soft-biometrics traits [61, 32, 17, 54, 52, 24, 25, 36, 29, 34, 31, 35] can be fused for improved gender classification, especially under more unconstrained scenarios.

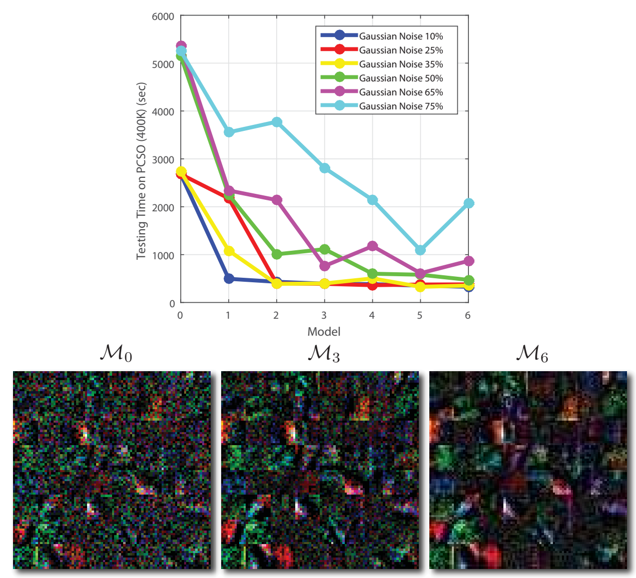


Figure 10: (Top) Testing time for the additive Gaussian noise occlusion experiments on various models. (Bottom) Visualization of the 64 first-layer filters for models $\mathcal{M}_0$, $\mathcal{M}_3$, and $\mathcal{M}_6$, respectively.

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 1

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine

translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[3] A. Bartle and J. Zheng. Gender classification with deep learning. 2

[4] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015. 3

[5] D.-Y. Chen and K.-Y. Lin. Robust gender recognition for uncontrolled environment of real-life images. *Consumer Electronics, IEEE Transactions on*, 56(3):1586–1592, 2010. 8

[6] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *IEEE CVPR*, pages 3286–3293, 2014. 2

[7] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *NIPS*, pages 577–585. 2015. 3

[8] W. Chuan-xu, L. Yun, and L. Zuo-yong. Algorithm research of face image gender classification based on 2-d gabor wavelet transform and svm. In *Computer Science and Computational Technology, 2008. ISCSCT'08. International Symposium on*, volume 1, pages 312–315. IEEE, 2008. 8

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, volume 1, page 2, 1990. 8

[10] S. Gutta, J. R. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *Neural Networks, IEEE Transactions on*, 11(4):948–960, 2000. 8

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 2

[12] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE TPAMI*, 34(1):194–201, 2012. 2

[13] S. Y. D. Hu, B. Jou, A. Jaech, and M. Savvides. Fusion of region-based representations for gender identification. In *IEEE/IAPR IJCB*, pages 1–7, Oct 2011. 2, 7, 8

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, (11):1254–1259, 1998. 2

[15] T. Jabid, M. H. Kabir, and O. Chae. Gender classification using local directional pattern (ldp). In *IEEE ICPR*, pages 2162–2165. IEEE, 2010. 8

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2008–2016, 2015. 2

[17] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, and M. Savvides. Gait-ID on the Move: Pace Independent Human Identification Using Cell Phone Accelerometer Dynamics. In *IEEE BTAS*, pages 8–15, Sept 2012. 8

[18] F. Juefei-Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides. Robust Local Binary Pattern Feature Sets for Periocular Biometric Identification. In *IEEE BTAS*, pages 1–8, sep 2010. 2

[19] F. Juefei-Xu, M. Cha, M. Savvides, S. Bedros, and J. Trojanova. Robust Periocular Biometric Recognition Using Multi-level Fusion of Various Local Feature Extraction Techniques. In *IEEE DSP*, pages 1–7, 2011. 2

[20] F. Juefei-Xu, K. Luu, and M. Savvides. Spartans: Single-sample Periocular-based Alignment-robust Recognition Technique Applied to Non-frontal Scenarios. *IEEE Trans. on Image Processing*, 24(12):4780–4795, Dec 2015. 2

[21] F. Juefei-Xu, K. Luu, M. Savvides, T. Bui, and C. Suen. Investigating Age Invariant Face Recognition Based on Periocular Biometrics. In *IEEE/IAPR IJCB*, pages 1–7, Oct 2011. 2

[22] F. Juefei-Xu, D. K. Pal, and M. Savvides. Hallucinating the Full Face from the Periocular Region via Dimensionally Weighted K-SVD. In *IEEE CVPRW*, pages 1–8, June 2014. 2

[23] F. Juefei-Xu, D. K. Pal, and M. Savvides. Methods and Software for Hallucinating Facial Features by Prioritizing Reconstruction Errors, 2014. U.S. Provisional Patent Application Serial No. 61/998,043, June 17, 2014. 2

[24] F. Juefei-Xu, D. K. Pal, and M. Savvides. NIR-VIS Heterogeneous Face Recognition via Cross-Spectral Joint Dictionary Learning and Reconstruction. In *IEEE CVPRW*, pages 141–150, June 2015. 8

[25] F. Juefei-Xu, D. K. Pal, K. Singh, and M. Savvides. A Preliminary Investigation on the Sensitivity of COTS Face Recognition Systems to Forensic Analyst-style Face Processing for Occlusions. In *IEEE CVPRW*, pages 25–33, June 2015. 8

[26] F. Juefei-Xu and M. Savvides. Can Your Eyebrows Tell Me Who You Are? In *IEEE ICSPCS*, pages 1–8, Dec 2011. 2

[27] F. Juefei-Xu and M. Savvides. Unconstrained Periocular Biometric Acquisition and Recognition Using COTS PTZ Camera for Uncooperative and Non-cooperative Subjects. In *IEEE WACV*, pages 201–208, Jan 2012. 2

[28] F. Juefei-Xu and M. Savvides. An Augmented Linear Discriminant Analysis Approach for Identifying Identical Twins with the Aid of Facial Asymmetry Features. In *IEEE CVPRW*, pages 56–63, June 2013. 2

[29] F. Juefei-Xu and M. Savvides. An Image Statistics Approach towards Efficient and Robust Refinement for Landmarks on Facial Boundary. In *IEEE BTAS*, pages 1–8, Sept 2013. 8

[30] F. Juefei-Xu and M. Savvides. Subspace Based Discrete Transform Encoded Local Binary Patterns Representations for Robust Periocular Matching on NIST's Face Recognition Grand Challenge. *IEEE Trans. on Image Processing*, 23(8):3490–3505, aug 2014. 2

[31] F. Juefei-Xu and M. Savvides. Encoding and Decoding Local Binary Patterns for Harsh Face Illumination Normalization. In *IEEE ICIP*, pages 3220–3224, Sept 2015. 8

[32] F. Juefei-Xu and M. Savvides. Facial Ethnic Appearance Synthesis. In *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 825–840. Springer International Publishing, 2015. 8

[33] F. Juefei-Xu and M. Savvides. Pareto-optimal Discriminant Analysis. In *IEEE ICIP*, pages 611–615, Sept 2015. 2

[34] F. Juefei-Xu and M. Savvides. Pokerface: Partial Order Keeping and Energy Repressing Method for Extreme Face Illumination Normalization. In *IEEE BTAS*, pages 1–8, Sept 2015. 8

[35] F. Juefei-Xu and M. Savvides. Single Face Image Super-Resolution via Solo Dictionary Learning. In *IEEE ICIP*, pages 2239–2243, Sept 2015. 8

[36] F. Juefei-Xu and M. Savvides. Weight-Optimal Local Binary Patterns. In *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 148–159. Springer International Publishing, 2015. 8

[37] F. Juefei-Xu and M. Savvides. Multi-class Fukunaga Koontz Discriminant Analysis for Enhanced Face Recognition. *Pattern Recognition*, 52:186–205, apr 2016. 2

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 3

[39] P.-H. Lee, J.-Y. Hung, and Y.-P. Hung. Automatic gender recognition using fusion of facial strips. In *IEEE ICPR*, pages 1140–1143. IEEE, 2010. 8

[40] X. Leng and Y. Wang. Improving generalization for gender classification. In *IEEE ICIP*, pages 1656–1659. IEEE, 2008. 8

[41] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE CVPRW*, June 2015. 2

[42] H. Lin, H. Lu, and L. Zhang. A new automatic recognition system of gender, age and ethnicity. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, volume 2, pages 9988–9991. IEEE, 2006. 8

[43] H. Lu and H. Lin. Gender recognition using adaboosted feature. In *Natural Computation, 2007. ICNC 2007. Third International Conference on*, volume 2, pages 646–650. IEEE, 2007. 8

[44] L. Lu, Z. Xu, and P. Shi. Gender classification of facial images based on multiple facial regions. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, volume 6, pages 48–52. IEEE, 2009. 8

[45] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sept 2015. 3

[46] A. Martinez and R. Benavente. The AR Face Database. *CVC Technical Report No.24*, June 1998. 4

[47] J. Merkow, B. Jou, and M. Savvides. An exploration of gender identification using only the periocular region. In *IEEE BTAS*, pages 1–5, Sept 2010. 2

[48] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014. 1

[49] B. Moghaddam and M.-H. Yang. Gender classification with support vector machines. In *IEEE FG*, pages 306–311. IEEE, 2000. 8

[50] D. K. Pal, F. Juefei-Xu, and M. Savvides. Discriminative Invariant Kernel Features: A Bells-and-Whistles-Free Approach to Unsupervised Face Recognition and Pose Estimation. In *IEEE CVPR*, June 2016. 2

[51] M. Savvides and F. Juefei-Xu. Image Matching Using Subspace-Based Discrete Transform Encoded Local Binary Patterns, Sept. 2013. US Patent US 2014/0212044 A1. 2

[52] K. Seshadri, F. Juefei-Xu, D. K. Pal, and M. Savvides. Driver Cell Phone Usage Detection on Strategic Highway Research Program (SHRP2) Face View Videos. In *IEEE CVPRW*, pages 35–43, June 2015. 8

[53] C. Tai, T. Xiao, X. Wang, and W. E. Convolutional neural networks with low-rank regularization. *ICLR*, abs/1511.06067, 2016. 3

[54] S. Venugopalan, F. Juefei-Xu, B. Cowley, and M. Savvides. Electromyograph and Keystroke Dynamics for Spoof-Resistant Biometric Authentication. In *IEEE CVPRW*, pages 109–118, June 2015. 8

[55] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE CVPR*, pages 3156–3164, June 2015. 3

[56] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE CVPR*, pages 842–850, June 2015. 3

[57] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv preprint arXiv:1511.05234*, 2015. 3

[58] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell - neural image caption generation with visual attention. In *ICML*, pages 2048–2057, June 2015. 1, 2

[59] Z. Yang, M. Li, and H. Ai. An experimental study on automatic face gender classification. In *IEEE ICPR*, volume 3, pages 1099–1102. IEEE, 2006. 8

[60] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *IEEE ICCV*, pages 4507–4515, Dec 2015. 3

[61] N. Zehngut, F. Juefei-Xu, R. Bardia, D. K. Pal, C. Bhagavatula, and M. Savvides. Investigating the Feasibility of Image-Based Nose Biometrics. In *IEEE ICIP*, pages 522–526, Sept 2015. 8

[62] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*, 2015. 3