Age and Gender Recognition in the Wild with Deep Attention

Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F.Xavier Roca, Jordi Gonzàlez

PII: S0031-3203(17)30254-6
DOI: 10.1016/j.patcog.2017.06.028
Reference: PR 6193

To appear in: *Pattern Recognition*

Received date: 2 February 2017
Revised date: 17 May 2017
Accepted date: 25 June 2017

Please cite this article as: Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F.Xavier Roca, Jordi Gonzàlez, Age and Gender Recognition in the Wild with Deep Attention, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.06.028

Highlights

- A novel feedforward attention mechanism for CNNs is proposed.
- The mechanism increases CNNs robustness to image deformations and clutter.
- The proposed mechanism increases CNNs performance for age and gender recognition.

Age and Gender Recognition in the Wild with Deep Attention

Pau Rodríguez[†], Guillem Cucurull[†], Josep M. Gonfaus[‡], F. Xavier Roca[†], Jordi Gonzàlez[†,‡]

[†] *Computer Vision Center and Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Catalonia Spain*

[‡] *Visual Tagging Services, Parc de Recerca, Campus UAB, 08193 Bellaterra, Catalonia Spain*

Abstract

Face analysis in images in the wild still pose a challenge for automatic age and gender recognition tasks, mainly due to their high variability in resolution, deformation, and occlusion. Although the performance has highly increased thanks to Convolutional Neural Networks (CNNs), it is still far from optimal when compared to other image recognition tasks, mainly because of the high sensitiveness of CNNs to facial variations. In this paper, inspired by biology and the recent success of attention mechanisms on visual question answering and fine-grained recognition, we propose a novel feedforward attention mechanism that is able to discover the most informative and reliable parts of a given face for improving age and gender classification. In particular, given a downsampled facial image, the proposed model is trained based on a novel end-to-end learning framework to extract the most discriminative patches from the original high-resolution image. Experimental validation on the standard Adience, Images of Groups, and MORPH II benchmarks show

that including attention mechanisms enhances the performance of CNNs in terms of robustness and accuracy.

Keywords: Age recognition, Gender recognition, Deep Neural Networks, Attention mechanisms

1. Introduction

Human face analysis constitutes one of the most important tasks in computer vision, since the automatic analysis of such a deformable object is of great importance [1]: the characterization of age, gender, facial attributes, expressions, garment, and even personality, to cite but a few, are crucial in several applications, like user identification, social interaction, face tracking, and behavior recognition [2, 3]. Regarding age and gender classification, although these two tasks have been largely addressed in the past, the reported performances are far from optimal [4, 5].

In the last few years, Convolutional Neural Networks (CNN) [6] have become the main workhorse for age and gender estimation. CNNs have been proven to perform very well in a variety of computer vision tasks such as human action recognition [7], handwritten digit recognition [8], face verification [9] or automatic face detection [10]. In relation to the task of soft-biometrics analysis, CNNs have been recently applied to the task of apparent age estimation [11, 12, 13], gender and smile classification [14], and real age and gender prediction [15]. However, due to the high variability of facial images in the wild, i.e. for example collected from the web, the low performance of CNNs

in tasks like age recognition shows that there is still room for improvement.

The main contribution of this paper is a novel feedforward attention mechanism that enhances current CNNs' robustness for highly variable unconstrained recognition tasks. Thus, inspired by biology, and the recent success of attention mechanisms [16], we propose a feedforward attention mechanism to discover the most discriminative patches of low resolution unconstrained facial images in order to process them in high resolution. So, beyond the increase in resolution, our method allows the network to assign more importance to the least occluded or deformed parts of the image, thus becoming the model more robust to noise and distractors. We perform a thorough evaluation on standard age and gender recognition benchmarks [17], proving that our attention pipeline is more robust than any previous state-of-the-art CNNs pretrained for facial recognition.

In particular, including attention on CNNs shows an increases in performance for standard CNNs such as VGG-16 [18] when applied to the Adience [17], Images of Groups (IoG) [19] and MORPH II datasets [20] for the tasks of age and gender recognition. Thus, on one hand, Adience and IoG consist of unconstrained facial images captured in the wild, showing that our model is capable of detecting soft biometric traits such as age and gender from facial pictures captured in uncontrolled environments, with distractions, deformations, and occlusions. Moreover, on the other hand, the proposed mechanism also shows improvement in controlled environments such as the MORPH II databaset, thanks to using higher resolution fixations.

2. Related Work

This section discusses other work that is relevant to understand our approach, together with the context and historical evolution in the use of neural networks for gender and age recognition.

2.1. Age Recognition

Not only the first studies in the 90s used the analysis of facial geometry [21] to estimate the age of a person, but also more recent techniques like the pipeline used in [22], presenting a combination of Biologically Inspired Features (BIF) and then using Canonical Correlation Analysis (CCA) and Partial Least Square (PLS) based methods. Indeed BIF were already used in [23] to represent face images, paving the way to works like [4, 24], showing that the automatic approach had matched the human performance. In fact, most of the approaches previous to CNNs were based on a two-stage pipeline, i.e. extracting features such as Local Binary Patterns (LBP) [17], and then classifying with a Support Vector Machine (SVM), or a Multi-layer Perceptron (MLP) [25, 26]. On the contrary, CNN based methods typically implements the two-step pipeline described above in just one step: the network learns both extracting the best features and either classifying such features into age categories [15, 27] or performing age regression [28, 29]. Deeper CNN models have been also applied to age and gender recognition [30], although most of them depend on domain-specific pre-training [31, 32]. Cascaded combinations of deep models were also considered in [33].

CNNs for facial images analysis have not been restricted to age estimation, but also to face verification, facial attribute estimation, and gender recognition. One illustrative example is the method presented in [34] which achieves a 99.2% face verification accuracy on the challenging Labeled Faces in the Wild dataset [35]. Unfortunately, this so impressive performance has not been yet achieved in other facial analysis tasks like gender recognition, for example, as shown next.

2.2. Gender Recognition

Regarding gender recognition, in contrast to age analysis, there is work from the early 90s where neural networks were already proposed, like the pioneering approach presented in [36]: authors proposed two neural network structures, an autoencoder and a classifier whose input was the encoded output layer of the autoencoder. The drawback of this method was that it relied on manual cropping, scaling and rotating the face of the picture, which was taken in a controlled environment.

Inspired from the age estimation methodology, pipelines based on a feature extractor and a stacked classifier were also proposed like in [37], [38], and [39]. On the other hand, the same CNN-based methods used for age were also applied to gender [28], [15], demonstrating that CNNs are truly capable of learning how to perform different tasks without any modification besides the data used for learning. For example, in [40] a CNN is trained to perform gender recognition by fine-tuning a pre-trained network, and then

an SVM is trained using the deep features computed by the CNN.

2.3. Neural Networks with Attention

Attention is a powerful mechanism that allows neural networks to look in more detail into particular regions of the input image to reduce the task complexity and discard irrelevant information, mildly inspired in the eye fixations performed by the human visual system [16].

Previous approaches for applying bioinspired visual attention mechanisms rely on finding visually salient regions on the image for processing them in a posterior step [41, 42]. In the context of neural networks, Larochelle and Hinton [43] proposed a third-order Restricted Boltzmann Machine (RBM) to combine high resolution "glimpses" of a sequence of fixations for image classification. Likewise, Denil *et al.* perform image tracking with an RBM fed with foveated images selected by a control pathway [44]. A simpler model was proposed by Ranzato [45], which predicts a glimpse location from a downsampled image and then uses it to extract a high-resolution patch. Spatial Transformer Networks (STN) [46] can be also considered as a form of attention, however, differently from other attention approaches like the one presented in this paper, they focus on a single spatially continuous region of the image. In all these proposed papers, attention is shown specially well-suited for images in the Wild, with multiple occlusions and distractors, which is the case of Adience and IoG datasets.

More recently, RNNs have become central to attention mechanisms since

they naturally integrate the information extracted from glimpses at different time-steps [16]. The ability to look "into the past" has made RNN-based attention mechanisms ideal for Natural Language Processing (NLP) tasks such as Neural Machine Translation [47], text-based question answering [48], image captioning [49], and Visual Question Answering (VQA) [50, 51].

In our work, we assume faces have already been detected, cropped, and aligned, and thus, there is no need to do a sequential search through the image with an RNN so as to find the most relevant image regions. However, since the main hypothesis of all the aforementioned papers is that CNN models can not give the same importance to all the regions of an image, mainly due to the high variability of unconstrained environments, attention mechanisms can be suitable in our case to automatically select specific regions of a face for further processing them in more detail, while ignoring background clutter. Based on these findings, we next describe our proposed attention-based CNN models.

3. Feedforward attention for age and gender recognition

The proposed model consists of three basic modules, see Figure 1: (i) an attention CNN ("where") that predicts the best attention grid to perform the glimpses, (ii) a patch CNN ("what") that evaluates the higher resolution patches based on their importance predicted by the attention grid, and (iii) a Multi Layer Perceptron (MLP) that integrates the information obtained from both CNNs and performs the final classification. We detail these mod-

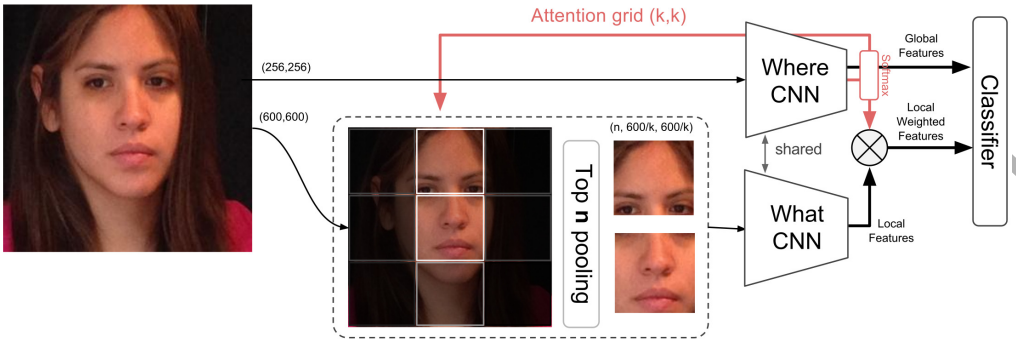


Figure 1: **The proposed attention model.** A lower resolution image is fed to the "where" CNN, which predicts a $k \times k$ attention grid. This grid is then used to extract high-resolution patches, from which the top n are pooled. The patches are then fed to the "what" CNN, whose output is weighted by the attention grid. Finally, feature maps from the "where" and "what" streams are concatenated and fed to an MLP classifier.

ules next.

The attention CNN is fed with all training images. We used the VGG-16 model since it has become the well-performing standard CNN that is supported in most of the deep learning programming frameworks [18], but any other CNN could be considered instead. This CNN is specifically trained to predict an attention $k \times k$ grid $\mathbf{G}$:

$$\mathbf{G}^{(k \times k)} \in \mathbb{R}_{\geq 0}, \qquad \sum_{i,j} \mathbf{G}_{i,j} = 1,$$

where k is an arbitrary number, and the values $\mathbf{G}_{i,j}$ represent the (normalized) importance of each patch. Then, a high-resolution version of the input image is divided in $k \times k$ patches, and fed to the patch CNN.

The patch CNN is fed with high resolution patches of the faces. Similarly to the attention CNN, any model could be used for this task, however, to reduce the computational requirements of this architecture, we reuse the first convolutional layers of the attention CNN. The output of this module consists of a matrix:

$$\mathbf{P}^{(k^2 \times d)} \in \mathbb{R},$$

where k^2 is the number of patches and d is the output dimension of the last convolutional layer. Global Average Pooling (GAP) is then used to reduce the spatial dimension of $\mathbf{P}$ to one, thus making it possible to feed images in their original resolution. These feature maps are subsequently weighted by $\mathbf{G}$ to reflect the importance of each patch of the grid.

In the literature, such weighting can be performed using either a weighted sum, denoted as "soft attention" in [49]:

$$\mathbf{P}^* = \mathbf{g} \cdot \mathbf{P}_{\mathbf{GAP}}, \tag{1}$$

or alternatively the element-wise product, also called the Hadamard product:

$$\mathbf{P}^* = \mathbf{g} \circ \mathbf{P}_{\mathbf{GAP}}; \tag{2}$$

where $\mathbf{g}^{(\mathbf{1} \times \mathbf{k^2})} \in \mathbb{R}_{\geq 0}$ is a flat view of $\mathbf{G}$. In the experimental section, we show that both strategies yield similar results. Then, on one hand, the Hadamard

product can be chosen to reduce the computational time complexity at the expense of memory. And, on the other hand, the weighted sum can be chosen in limited memory scenarios but conveying higher computational time cost.

Additionally, for further reducing the computational cost, we pool the top n patches before feeding them to the patch CNN, thus using a "hard attention" mechanism instead. It is important to note that the gradients will not propagate to those grid positions outside the top n, but since the importance given to those discarded positions are zero or very close to zero, the network is still able to learn. Additionally, as it is usually done in the literature, we also performed random patch sampling given the distribution of the attention grid, however. the difference in performance when compared to sampling the top n patches is not statistically significant.

The classifier is fed with features from the `pool5` layer of the attention CNN, and the weighted features of either the `pool4` or `pool3` layers of the patch CNN. Lower level features maps from the patch CNN are preferred because they correspond to local-level image features.

We also consider two strategies for merging the feature maps of both CNNs: (i) concatenate them after an L2 normalization, and (ii) learn a projection of the patch CNN feature maps to the attention CNN feature map space, and simply add them. In the next section, we demonstrate that the normed concatenation yielded slightly better results than the project-and-add strategy.

Figure 2: **Adience sample.** Sample of each age group and gender from the fourth fold of the Adience dataset.

The resulting feature maps are then fed to the final classifier, which consists of the `fc6`, `fc7`, and `fc8` layers of the VGG-16, as typically done in the CNN literature. In the following sections, several experiments are presented for testing the robustness and accuracy of the whole attention-based architecture.

4. Benchmark Datasets

To evaluate the performance of our approach on unconstrained facial images, we test it on the Adience dataset proposed in [17], and following the same evaluation benchmark. This dataset consists of 26.5K images distributed in eight age categories (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+) with the corresponding gender label.

The Adience benchmark measures both the accuracy in gender and age recognition tasks using 5-fold cross-validation in which the provided folds are subject-exclusive. The final score is given by the mean of the accuracies of the five folds. The same subject-exclusive folds are used for age and gender,

Reference	**Features**	**Classifier**	**Task**	**Accuracy** (%)
Levi [15]	Learned	CNN	Age	50.7
Chen [33]	Learned	CNN	Age	52.9
Rothe [32]	Learned	CNN	Age	55.6
Ozbulak [31]	Learned	CNN	Age	57.9
Rothe [32]	Learned	CNN	Age	64.0
Eidinger [17]	LBP+FPLBP	SVM	Gender	76.1
Tapia [52]	LBP	SVM	Gender	79.8
Levi [15]	Learned	CNN	Gender	86.8
Wolfshaar [40]	Learned	CNN	Gender	87.2
Ozbulak [31]	Learned	CNN	Gender	92.0

Table 1: **Previous results on the Adience dataset.**

and there is a subset of only nearly frontal faces which we have not used since faces in real world images present a higher diversity on the pose. This dataset is designed to be as similar as possible as real-world challenging face images, therefore, the faces present many changes in pose, rotation, appearance, light and noise. Figure 2 shows some samples of the dataset presenting significant differences between them. It is important to remark that the eight age range groups and the number of samples per age class are not equally distributed since some classes have more samples than the others.

In Table 1 the previous results published on the Adience dataset for age and gender estimation are listed. The results can be divided between the ones using LBP features (and variants) and the ones using a deep learning approach. As expected, CNNs yield significantly better results than SVMs on LBP features extracted from facial images.

To test the generalization capability of our models on age recognition, we

Adience [17]	0-2	4-6	8-13	15-20	25-32	{38-43, 48-53}	60+
IoG [19]	0-2	3-7	8-12	13-19	20-36	37-65	66+

Table 2: **Mapping between Adience and IoG.** This Table shows the mapping between Adience and Images of Groups age categories to perform cross-dataset evaluation.

have tested them using the Images of Groups (IoG) dataset presented in [19]. This dataset consists of $5.1K$ images of groups of people where $28.2K$ faces have been annotated with gender and age group labels. Like the images from the Adience dataset, the ones from IoG dataset present several differences in pose, appearance, and light, and they are even more challenging because the size of the faces is much smaller than Adience faces.

The 7 age groups from this dataset are quite similar to the 8 groups used in Adience dataset, so we can train models on the Adience dataset and then evaluate them on the IoG dataset: the mapping between Adience and IoG age categories is defined in Table 2.

In Table 3 there is a listing of all the previous results on the Images of groups dataset, indicating the methods used to tackle the problems of age or gender classification.

Deep Learning was also used in this dataset by Mansanet et al. [30] and Dong et al. [60], achieving good results in both age and gender recognition. In these deep learning approaches, similarly to our proposed network, the best features to perform age or gender recognition are not hand-crafted but learned from the data.

In order to evaluate the advantages of the proposed mechanism in a con-

Ref.	Features	Classifier	Task	Data Split	Acc. (%)
[53]	ML-LPQ	SVM	age	original	56.0
[54]	OHLG	SVM	age	custom1	59.5
[55]	LBP+SIFT+CH	SVM	age	Dago's [56]	63.0
[17]	LBP+FPLBP	SVM	age	Dago's [56]	66.6
[39]	BIF	SVM	age	5 fold	68.1
[57]	HOG+LBP+LTP+WLD	SVM	gender	Dago's [56]	92.5
[58]	ASR+	SRC	gender	custom3	93.3
[52]	LBP	SVM	gender	custom3	94.6
[30]	Local-DNN	MLP	gender	Dago's [56]	96.3
[59]	CNN+HOG+LBP+LOSIB	SVM+CNN	gender	Dago's [56]	97.2

Table 3: **Previous results on the Images of Groups Dataset.** [1], [2], [3]: Different data splits used by the authors.

Reference	**Method**	**Classifier**	**Task**	**MAE**
Chang [61]	AAM	OHRank	age	5.69
Wang [27]	CNN	DLA	age	4.77
Rothe [62]	CNN	SVR	age	3.45
Huerta [29]	CNN	MLP	age	3.31
Rothe [32]	CNN	DEX	age	2.68

Table 4: **Previous results on the MORPH II dataset** using the same data split. For the sake of clarity and fairness in the comparison, all these reported MAEs use the same data split and data subsets, only [29] provide a more complete evaluation procedure, using different data splits of 44K (MAE 3.31) and 55K (MAE 3.88).

trolled environment (i.e. centered, unoccluded faces with common background), we test it on the MORPH II dataset, which consists of more than 50K mug shots. We follow a well-known experimental setup in the literature [63, 61, 11, 32] consisting of a subset of 5474 pictures with ages comprised between 16 and 77 years old. From the subset, a 80% of them for training and a 20% for validation. The performance of previous approaches on the

same data split of the MORPH II dataset are listed in Table 4. Performance is reported in Mean Average Error (MAE), the standard error measure for age regression in the literature:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n},\ \{Y, \hat{Y}\} \in \mathbb{R} \tag{3}$$

where y_i is the ground truth value corresponding to the ith example, and $\hat{y}_i$ is the predicted value for that example.

5. Experimental Evaluation

Our model is based on the VGG-16 CNN [18], and it is implemented with Tensorflow [64]. We use domain-specific pre-training for initializing the CNN weights since it has been proved to achieve better performance than pre-training in general tasks such as Imagenet [31, 32]. Parameters are initialized with the standard VGG-16 architecture trained for face recognition on 2.6M images of 2.6K people [65] since, unlike [11], (i) it has been also tested for gender recognition, (ii) it uses the base VGG-16 model (without DEX), (iii) it focuses in a higher variety of facial analysis tasks than IMDB-WIKI, and (iv) it would deviate from the main purpose of this work, which is to evaluate the effects of attention mechanisms of CNNs for facial recognition tasks. The fully-connected layers are initialized with the Xavier initializer [66]. Models are optimized with `sgd` for 30 epochs, or until they reach a plateau. The learning rate is initially set to 0.0001 and divided by 10 every ten epochs. All

		Accuracy (%)
Baseline VGG-16 (no Attention)		57.80
VGG-16 + SVM [31]		57.90
Attention mode	No grid	59.41
	Project.	61.42
	Eltwise	61.78
Weight Sharing	No	61.37
	Yes	61.55
Merge mode	Add	61.35
	Concat.	61.78

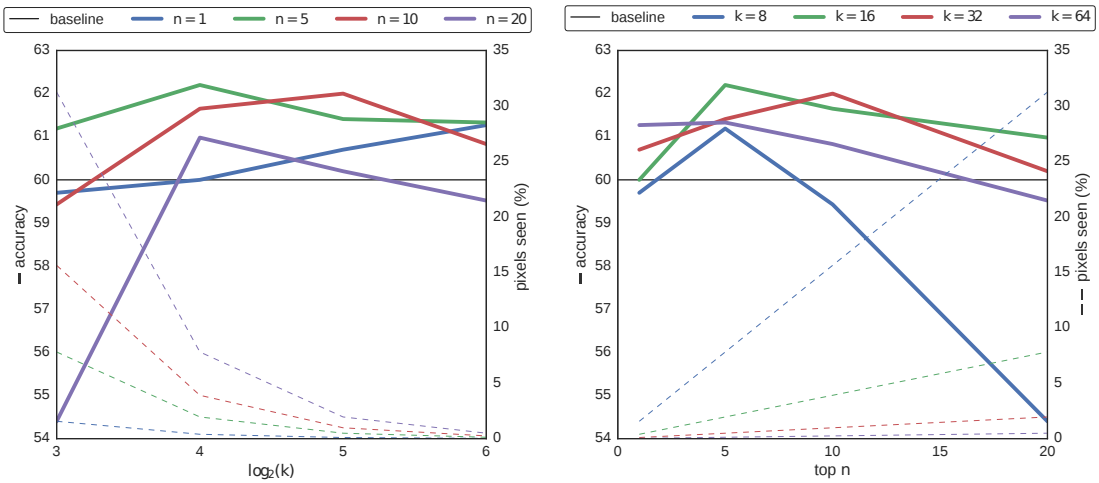
Table 5: **Average validation error rate** when fixating different properties of the attention mechanism and random combinations of the rest, on a random fold of the Adience dataset. The proposed attention mechanism is proven robust to the weight sharing, different strategies of weighting the patches, and merging the feature maps from the attention and patch streams.

the other hyper-parameters are found by random search unless we explicitly specify otherwise.

Next, we evaluate the influence of the different design decisions in the proposed model, namely the attention mode, weight sharing, merge mode, attention grid and patch depth. Following the same procedure as in the related works applied to this dataset, any possible design decision is firstly evaluated on a random fold to make the experimentation tractable, since results are proven consistent between folds.

Attention mode. We found that performing the Hadamard product between the attention weights and the patch feature maps was slightly better than the weighted sum ("Project." in Table 5). No weighting at all (no attention) resulted in the worst performance.

Weight sharing. As it can be seen in Table 5, using independent weights



(a) Fixate top-n (b) Fixate grid k

Figure 3: **Maximum accuracy for different combinations of grid size** (k), and choosing n patches. Dashed lines represent the accumulated percentage of patch pixels with respect to the whole image. Dividing the image in 16 regions, and choosing the top-5 patches results in best performance by just processing a 2% of the high-resolution pixels.

for the patch network yields small improvements in average. However, we observed the opposite effect for big patch network inputs due to overfitting, *e.g.* $n = 20$, $k = 8$. Thus, since the proposed model is robust to changes in weight sharing, we decided to keep the weights shared to reduce the memory consumption of the network.

Merge mode. Compared to projecting the patch feature maps to the attention stream output space and adding them, feeding the classifier with the normed concatenation both streams resulted in the best performance by a small margin.

The Attention Grid. As it was shown in Figure 1, a $k \times k$ weight grid is predicted on the low-resolution input image. Since this grid is used to extract

Figure 4: **The predicted attention grid.** The top row corresponds to a grid learned with $k = 4, n = 4$, and the bottom row to $k = 16, n = 5$. High attention is shown in red, and low attention in blue. As it can be seen, the attention grid predicts low values for background, glasses, and rings, while it is most centered

n patches, we can control the portion of the image that will be fed to the patch CNN by tuning n and k. Figure 3 shows the impact on performance of choosing different combinations of k, and n. As it can be seen, for most combinations, our approach outperforms the baseline score by a 2.4% margin with $n = 5$, and $k = 16$. Samples of the attention grid are shown in Figure 4, which corresponds to the predicted attention grid for $k = 4, n = 4$, and $k = 16, n = 5$.

Patch CNN Depth. Given that the convolutional layers of the attention path are reused, the depth of the patch CNN is conditioned to the five convolutional modules of the VGG-16. From this CNN, we compared features from the `pool3`, and `pool4` because they shrink the input size, and they are less invariant than features from `pool5`. We empirically found that using

`pool3` yielded a 0.6% improvement over `pool4`. And this is consistent with the fact that lowest level features maps from the patch CNN are preferred since they better correspond to local image features.

Summarizing, including attention in CNN models is robust to any possible design configurations as presented in this paper, *i.e.* the attention grid application mode, the feature merging, and weight sharing. In fact, the most critical hyperparameters are the grid size k, and the number n of patches to feed to the patch network. For the next section, the parameters were fixed to `attention mode=eltwise, weight sharing=yes, merge mode=concat, k=16, n=5`, and the high-resolution images size is 600×600 before random cropping and random flip.

6. Evaluation on age and gender recognition

As it can be seen in Table 6, implementing the proposed attention mechanism on VGG-Faces [65] increases the accuracy in 4% on age recognition and 0.6% on gender recognition when not considering attention [31]. For the age classification problem, 1-off accuracy is also reported, indicating the accuracy of our model considering a one error distance prediction as correct. As expected, the 1-off accuracy of our model is 2.3% higher than the best-reported accuracy with VGG-16 pretrained with Faces.

To the best of our knowledge, the top accuracy score obtained in gender classification Adience benchmark is 92.0% [31], and their results on age esti-

Model	Accuracy (%) Age	1-off	Gender
Eidinger [17]	45.1	80.7	77.8
Tapia [52]	-	-	79.8
Levi [15]	50.7	84.7	86.8
Wolfshaar [40]	-	-	87.2
Chen [33]	52.9	-	-
Rothe (VGG16-DEX) [32]	55.6	89.7	-
Ozbulak (VGG16-Faces + SVM) [31]	57.9	-	92.0
Rothe (VGG16-DEX-IMDB) [32]	**64.0**	**96.6**	-
Ours			
VGG16-Faces	57.8 ± 4.9	92.8 ± 1.8	92.4 ± 1.9
VGG16-Faces + Attention	61.8 ± 2.1	95.1 ± 0.03	**93.0** ± 1.8

Table 6: **Accuracies obtained on the Adience dataset for the 5 folds.** The VGG-16 model pre-trained on $> 3M$ faces [65] obtains the best performance when the attention mechanism is included.

		Predicted 0-2	4-6	8-12	15-20	25-32	38-43	48-53	+60
Real	**0-2**	**64.2**	34.8	0.6	0.1	0.1	0.0	0.1	0.1
	4-6	15.8	**70.1**	12.2	1.3	0.3	0.1	0.1	0.1
	8-12	1.9	17.2	**59.1**	14.9	5.8	0.6	0.4	0.1
	15-20	0.1	1.0	12.8	**40.7**	41.1	4.2	0.2	0.1
	25-32	0.0	0.3	1.7	11.4	**69.2**	16.3	1.0	0.0
	38-43	0.0	0.1	0.5	2.1	39.8	**47.4**	8.0	2.2
	48-53	0.0	0.0	0.2	0.8	7.8	41.6	**31.1**	18.5
	+60	0.1	0.0	0.2	0.2	4.1	14.8	27.0	**53.6**

Table 7: **Age confusion matrix for the Adience dataset**. This Table represents the confusion matrix of our model predictions over the whole dataset.

mation are 57.9%. When our proposed method is trained to perform gender recognition it also achieves state-of-the-art performance on facial gender classification. In contrast, [32] do not apply their approach to gender analysis.

Model name	Accuracy (%) Age	1-off	Gender
Gallagher [19]	42.9	78.1	74.1
Li [67]	48.5	88.0	-
Shan [68]	50.3	87.1	74.9
Ylioinas [69]	51.7	88.7	-
Dong [60]	54.0	91.0	-
Bekhouche [53]	56.0	88.8	79.1
VGG-16 (Adience) + Attention	**60.0**	**94.5**	**86.9**

Table 8: **Accuracies obtained on the Images of Groups dataset.** This Table shows the accuracies obtained on the IoG dataset by averaging the predictions of the models trained on the Adience dataset. Our results are compared with the previous work where the same age balanced test set has been used, as proposed in [19].

Reference	Method	Classifier	Task	MAE
Rothe [62]	CNN	SVR	age	3.45
Rothe [32]	CNN	DEX	age	2.68
VGG-16 [32] + Attention	CNN	DEX	age	**2.56**

Table 9: **MAE on the MORPH II dataset.** Adding attention to [32] decreases the MAE.

In order to evaluate how well the proposed attention model generalizes, a cross-dataset experiment was performed on IoG, see Table 8. The 5 models trained on the Adience dataset for gender recognition were used to classify the 1,050 test images from the IoG dataset and their predictions were averaged. As shown in Table 8, this ensemble obtained a gender classification accuracy of 86.9% on the IoG dataset, surpassing the state-of-the-art score from [53], which is 79.1% for this test split, thus confirming the generality of our approach. Table 9, shows the results on the MORPH II dataset. As it

	0-2	**4-6**	**8-12**	**15-20**	**25-32**	**38-43**	**48-53**	**+60**
Acc.(%)	84.6	82.2	89.0	96.1	98.2	98.6	97.3	94.2

(a) Adience dataset.

	0-2	**3-7**	**8-12**	**13-19**	**20-36**	**37-65**	**+66**
Acc.(%)	67.3	82.0	82.7	91.3	96.0	98.0	90.6

(b) IoG dataset.

Table 10: **Gender accuracy** per age for the Adience and IoG datasets.

can be seen, adding attention results in 2.56 MAE, a relative 4.47% improvement with respect to the state of the art [32]. Additionally, in Table 10, it is shown that the proposed approach performs very well on adults, whereas it fails more frequently when classifying very young subjects. This performance is expected as even for humans estimating the gender of young children is harder than the gender of adults.

7. Discussion

A novel feedforward CNN pipeline which incorporates an attention mechanism for automatic age and gender recognition for face analysis has been proposed. The presented model consists of an attention network which estimates the most informative patches in the low-resolution image, which are further processed in a patch network in higher resolution. As a result, the attention-based CNN is proven to be more robust to clutter and deformation, inherent in deformable objects like faces. Alternative design choices for implementing the attention pipeline (i.e. attention mode, weight sharing,

merge mode, attention grid, and patch network depth) have been proposed and compared, thus proving the robustness of the whole approach and consistently outperforming the model without attention.

Experiments show that networks enhanced with the proposed mechanism are more robust in in-the-wild tasks such as age and gender recognition in the Adience and IoG datasets. Concretely, enhanced models experienced a relative improvement of 8.75% for age recognition and a 7.89% on age classification with the Adience benchmark. The generality of the proposed model has also been demonstrated by performing a cross-dataset experiment, resulting in state-of-the-art performance on the IoG dataset. Moreover, experiments on MORPH II demonstrate that the proposed model enhances CNNs even in constrained environments with centered faces and gray backgrounds, resulting in a 4.47% relative improvement with respect to a state-of-the-art model [32]. An explanation for this effect is that the enhanced CNN has the ability to perform detailed fixations in the most discriminative patches depending on the context (for instance gender).

Qualitative results are shown in Figure 6, where images wrongly classified by VGG-16 (pre-trained on faces) are correctly classified by the proposed attention model. Also, it is shown how the attention mechanism is able to ignore clutter. Extreme rotations and occluding attributes (like fancy dressings) are the main reason of misclassifications, together with the presence of multiple people of different ages in the same image, or simply people who seem younger or older than their real age. This rises the interesting problem

Figure 5: **Gender misclassifications.** This figure shows several subjects whose gender have been misclassified. The first row contains females that were wrongly classified as males, whereas the second row contains males that were misclassified as females.

Figure 6: Corrected miss-classifications when adding our attention mechanism to VGG-16 pre-trained with faces [65]. Each row corresponds to an age group. Note that our approach is more robust to clutter.

of apparent age estimation, as recently addressed in [70].

For the case of gender recognition, the proposed model mostly fails with the youngest ages, difficult to be distinguished even by humans, see Fig. 5.

Acknowledgments

Authors acknowledge the support of the Spanish project TIN2015-65464-R (MINECO/FEDER), the 2016FI_B 01163 grant by the CERCA Programme - Generalitat de Catalunya, and the COST Action IC1307 iV&L Net (European Network on Integrating Vision and Language) supported by COST (European Cooperation in Science and Technology). We also gratefully ac-

knowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU and a GTX TITAN GPU, used for this research.

References

[1] A. Dantcheva, P. Elia, A. Ross, What else does your biometric data reveal? a survey on soft biometrics, Information Forensics and Security, IEEE Transactions on 11 (2015) 441–467.

[2] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, A. K. Jain, Unconstrained face recognition: Identifying a person of interest from a media collection, Information Forensics and Security, IEEE Transactions on 9 (2014) 2144–2157.

[3] J. Orozco, O. Rudovic, J. Gonzàlez, M. Pantic, Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises, Image and Vision Computing 32 (2014) 14–26.

[4] H. Han, C. Otto, A. K. Jain, Age estimation from face images: Human vs. machine performance, in: Biometrics (ICB), 2013 International Conference on, IEEE, 2013, pp. 1–8.

[5] J. E. Tapia, C. A. Perez, Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape, Information Forensics and Security, IEEE Transactions on 8 (2013) 488–499.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (1989) 541–551.

[7] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, TPAMI 35 (2013) 221–231.

[8] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: CVPR, IEEE, 2012, pp. 3642–3649.

[9] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: CVPR, IEEE, 2014, pp. 1701–1708.

[10] S. S. Farfade, M. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, arXiv preprint arXiv:1502.02766 (2015).

[11] R. Rothe, R. Timofte, L. Van Gool, Dex: Deep expectation of apparent age from a single image, in: ICCV Workshops, 2015, pp. 10–15.

[12] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, X. Chen, Agenet: Deeply learned regressor and classifier for robust apparent age estimation, in: ICCV Workshops, 2015, pp. 16–24.

[13] Z. Kuang, C. Huang, W. Zhang, Deeply learned rich coding for cross-dataset facial age estimation, in: ICCV Workshops, 2015, pp. 96–101.

[14] K. Zhang, L. Tan, Z. Li, Y. Qiao, Gender and smile classification using deep convolutional neural networks, in: CVPR Workshops, 2016, pp. 34–38.

[15] G. Levi, T. Hassncer, Age and gender classification using convolutional neural networks, in: CVPR Workshops, 2015, pp. 34–42.

[16] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: NIPS, 2014, pp. 2204–2212.

[17] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, Information Forensics and Security, IEEE Transactions on 9 (2014) 2170–2179.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CoRR abs/1409.4842 (2014).

[19] A. Gallagher, T. Chen, Understanding images of groups of people, in: Proc. CVPR, 2009.

[20] K. Ricanek, T. Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, 2006, pp. 341–345.

[21] Y. H. Kwon, N. D. V. Lobo, Age classification from facial images, in: CVPR, IEEE, 1994, pp. 762–767.

[22] G. Guo, G. Mu, Joint estimation of age, gender and ethnicity: Cca vs. pls, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–6.

[23] G. Guo, G. Mu, Y. Fu, C. Dyer, T. Huang, A study on automatic age estimation using a large database, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1986–1991.

[24] H. Han, C. Otto, X. Liu, A. K. Jain, Demographic estimation from face images: Human vs. machine performance, TPAMI 37 (2015) 1148–1161.

[25] T. Kanno, M. Akiba, Y. Teramachi, H. Nagahashi, A. Takeshi, Classification of age group based on facial images of young males by using neural networks, IEICE TRANSACTIONS on Information and Systems 84 (2001) 1094–1101.

[26] K. B. R. K Ramesha, Feature extraction based face recognition, gender and age classification 2 (2010).

[27] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, in: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, IEEE, 2015, pp. 534–541.

[28] D. Yi, Z. Lei, S. Z. Li, Age estimation by multi-scale convolutional network, in: Computer Vision–ACCV 2014, Springer, 2015, pp. 144–158.

[29] I. Huerta, C. Fernández, C. Segura, J. Hernando, A. Prati, A deep analysis on age estimation, Pattern Recognition Letters 68 (2015) 239–249.

[30] J. Mansanet, A. Albiol, R. Paredes, Local deep neural networks for gender recognition, Pattern Recognition Letters 70 (2016) 80–86.

[31] G. Ozbulak, Y. Aytar, H. K. Ekenel, How transferable are cnn-based features for age and gender classification?, in: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE, 2016, pp. 1–6.

[32] R. Rothe, R. Timofte, L. Van Gool, Deep expectation of real and apparent age from a single image without facial landmarks, International Journal of Computer Vision (2016) 1–14.

[33] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, R. Chellappa, A cascaded convolutional neural network for age estimation of unconstrained faces, in: Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on, IEEE, 2016, pp. 1–8.

[34] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, CoRR abs/1406.4773 (2014).

[35] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained

Environments, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[36] B. A. Golomb, D. T. Lawrence, T. J. Sejnowski, Sexnet: A neural network identifies sex from human faces., in: NIPS, volume 1, 1990, p. 2.

[37] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, A. M. Mirza, Gender recognition from face images with local wld descriptor, in: Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on, IEEE, 2012, pp. 417–420.

[38] C. Shan, Learning local binary patterns for gender classification on real-world face images, Pattern Recognition Letters 33 (2012) 431–437.

[39] H. Han, A. Jain, Age, gender and race estimation from unconstrained face images, Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

[40] J. van de Wolfshaar, M. F. Karaaba, M. A. Wiering, Deep convolutional neural networks and support vector machines for gender recognition (2015).

[41] L. Itti, C. Koch, E. Niebur, et al., A model of saliency-based visual attention for rapid scene analysis, TPAMI 20 (1998) 1254–1259.

[42] A. Torralba, A. Oliva, M. S. Castelhano, J. M. Henderson, Contextual

guidance of eye movements and attention in real-world scenes: the role of global features in object search., Psychological review 113 (2006) 766.

[43] H. Larochelle, G. E. Hinton, Learning to combine foveal glimpses with a third-order boltzmann machine, in: NIPS, 2010, pp. 1243–1251.

[44] M. Denil, L. Bazzani, H. Larochelle, N. de Freitas, Learning where to attend with deep architectures for image tracking, Neural computation 24 (2012) 2151–2184.

[45] M. Connolly, N. Jones, D. Turner, E-learning: a fresh look, Higher Education Management and Policy 18 (2006) 135.

[46] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: NIPS, 2015, pp. 2017–2025.

[47] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[48] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: NIPS, 2015, pp. 1693–1701.

[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, preprint arXiv:1502.03044 (2015).

[50] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, arXiv preprint arXiv:1511.02274 (2015).

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

[52] J. E. Tapia, C. A. Perez, Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape, IEEE Transactions on Information Forensics and Security 8 (2013) 488–499.

[53] S. E. Bekhouche, A. Ouafi, A. Benlamoudi, A. Taleb-Ahmed, A. Hadid, Facial age estimation and gender classification using multi level local phase quantization, in: Control, Engineering & Information Technology (CEIT), 2015 3rd International Conference on, IEEE, 2015, pp. 1–4.

[54] F. Alnajar, C. Shan, T. Gevers, J.-M. Geusebroek, Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions, Image and Vision Computing 30 (2012) 946–953.

[55] E. Fazl-Ersi, M. E. Mousa-Pasandi, R. Laganiere, M. Awad, Age and gender recognition using informative features of various types, in: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 5891–5895.

[56] P. Dago-Casas, D. González-Jiménez, L. L. Yu, J. L. Alba-Castro, Single-and cross-database benchmarks for gender classification under unconstrained settings, in: ICCV Workshops, 2011, pp. 2152–2159.

[57] M. Castrillón-Santana, J. Lorenzo-Navarro, E. Ramón-Balmaseda, On using periocular biometric for gender classification in the wild, Pattern Recognition Letters (2015).

[58] D. Mery, K. Bowyer, Recognition of facial attributes using adaptive sparse representations of random patches, in: ECCV Workshops, Springer, 2014, pp. 778–792.

[59] M. Castrillón-Santana, J. Lorenzo-Navarro, E. Ramón-Balmaseda, Descriptors and regions of interest fusion for gender classification in the wild, arXiv preprint arXiv:1507.06838 (2015).

[60] Y. Dong, Y. Liu, S. Lian, Automatic age estimation based on deep learning algorithm, Neurocomputing (2015).

[61] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: CVPR, 2011, pp. 585–592.

[62] R. Rothe, R. Timofte, L. Van Gool, Some like it hot-visual guidance for preference prediction, in: CVPR, 2016, pp. 5553–5561.

[63] G. Guo, Y. Fu, C. R. Dyer, T. S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, Image Processing, IEEE Transactions on 17 (2008) 1178–1188.

[64] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467 (2016).

[65] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC, 2015.

[66] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks., in: Aistats, volume 9, 2010, pp. 249–256.

[67] C. Li, Q. Liu, J. Liu, H. Lu, Learning ordinal discriminative features for age estimation, in: CVPR, IEEE, 2012, pp. 2570–2577.

[68] C. Shan, Learning local features for age estimation on real-life faces, in: Workshop on Multimodal Pervasive Video Analysis, ACM, 2010, pp. 23–28. doi:10.1145/1878039.1878045.

[69] J. Ylioinas, A. Hadid, M. Pietikainen, Age classification in unconstrained conditions using lbp variants, in: ICPR, 2012, pp. 1257–1260.

[70] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, I. Guyon, Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results, in: ICCV Workshops, 2015, pp. 1–9.

Pau Rodríguez received the MSc degree in Artificial Intelligence from KU Leuven in 2015. He is currently a Ph.D. student at the Image Social Evaluation (ISE) Lab of the Computer Vision Center and the Universitat Autònoma de Barcelona, Catalonia Spain. His research interests focus on machine learning, pattern recognition, and computer vision.

Guillem Cucurull received the BSc degree in Computer Science from Universitat Autònoma de Barcelona (UAB) in 2016. He is currently a MSc student in Computer Vision and a research assistant at the the Image Sequence Evaluation (ISE Lab) research group of the Computer Vision Center and UAB. His research interests include computer vision and machine learning techniques for human motion and behaviour analysis.

Josep M. Gonfaus received the PhD degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB) in 2012. He participated in various Pascal challenges. He co-founded a spin-off based on their research by applying computer vision and deep learning techniques for analyzing social media data (Visual Tagging). His main research interests are deep learning techniques to mimic human cognitive capabilities.

F. Xavier Roca received the Ph.D. degree in computer science from Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain, in 1990. He is an Associate Professor and the Director of the Department

of Computer Science, UAB. He is also a Research Fellow with the Computer Vision Center. He has been a Principal Researcher in several projects (public and private funds). He is working in technological transfer computer vision. The topics of his research are active vision, biometrics and tracking.

Jordi Gonzàlez received the PhD degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB) in 2004. He is an associate professor in computer science at the Computer Science Department, UAB. He is also a research fellow at the Computer Vision Center, where he has co-founded three spin-offs (Cloud Size Services, Visual Tagging, Care Respite) and the Image Sequence Evaluation (ISE Lab) research group. His research interests include machine learning techniques for the computational interpretation of social images, or Visual Hermeneutics.