Predicting Personal Traits from Facial Images using Convolutional Neural Networks Augmented with Facial Landmark Information

Yoad Lewenberg[1] yoadlew@cs.huji.ac.il

Yoram Bachrach[2] yobach@microsoft.com

Sukrit Shankar[3] ss965@cam.ac.uk

Antonio Criminisi[2] antcrim@microsoft.com

[1]The Hebrew University of Jerusalem, Israel
[2]Microsoft Research, Cambridge, United Kingdom
[3] Machine Intelligence Lab (MIL), Cambridge University

Abstract

We consider the task of predicting various traits of a person given an image of their face. We estimate both objective traits, such as gender, ethnicity and hair-color; as well as subjective traits, such as the emotion a person expresses or whether he is humorous or attractive. For sizeable experimentation, we contribute a new Face Attributes Dataset (FAD), having roughly 200,000 attribute labels for the above traits, for over 10,000 facial images.

Due to the recent surge of research on Deep Convolutional Neural Networks (CNNs), we begin by using a CNN architecture for estimating facial attributes and show that they indeed provide an impressive baseline performance. To further improve performance, we propose a novel approach that incorporates facial landmark information for input images as an additional channel, helping the CNN learn better attribute-specific features so that the landmarks across various training images hold correspondence. We empirically analyse the performance of our method, showing consistent improvement over the baseline across traits.

1 Introduction

Humans find it very easy to determine various traits of other people, simply by looking at them. Without almost any conscious effort, a glimpse at another person's face is sufficient for us to ascertain their gender, age or ethnicity. We can easily decide whether they are attractive, look funny or are approachable, or determine the emotion they are displaying (for example, whether they appear sad, happy or surprised). As social creatures, making such inference is clearly important to us. Imparting commensurate capabilities to machines is bound to enable very interesting applications [Parikh *et al.*, 2012]. However, in contrast to the relative ease with which humans infer such personal traits of an individual from their facial image, training a machine to do the same is a challenging task.

Deep Convolutional Neural Networks (CNNs) [LeCun *et al.*, 1998; Behnke, 2003; Simard *et al.*, 2003] are prominent statistical learning models, which have recently been shown to be very effective for image classification tasks [Krizhevsky *et al.*, 2012; Bengio *et al.*, 2007; Deng *et al.*, 2009; Szegedy *et al.*, 2014]. These networks employ several layers of neuron collections in a feed-forward manner, where the individual neurons are tiled in a way so that they respond to overlapping regions in the visual field. As opposed to hand crafted convolution kernel methods [Maini and Aggarwal, 2009], the elements of each convolution kernel in CNNs are trained by backpropagation, applied in conjunction with an optimization technique such as Stochastic Gradient Descent (SGD) [LeCun *et al.*, 1998].

Analyzing facial images has been a key research area in computer vision and artificial intelligence for quite a long time. Researchers have proposed automated methods for inferring personal traits of individuals from facial images [Lyons *et al.*, 1999], including gender [Moghaddam and Yang, 2002], age [Horng *et al.*, 2001] and ethnicity [Lu and Jain, 2004; Hosoi *et al.*, 2004].

Earlier work has even uncovered methods for predicting more subjective or social traits from facial images [Vinciarelli *et al.*, 2009], such as the expressed emotion [Padgett and Cottrell, 1997; Fasel and Luettin, 2003] or attractiveness [Kagian *et al.*, 2006; Datta *et al.*, 2008]. Very recently, Microsoft [2015] released mobile and web applications, which were aimed at guessing the age of humans by just looking at their facial images.

In addition to specific methods for predicting personal attributes, earlier works have also examined reusable building blocks for facial image processing, for tasks such as detecting faces, pose estimation, face segmentation and facial landmark localization [Huang *et al.*, 2007; Segundo *et al.*, 2010; Zhu and Ramanan, 2012; Uřičář *et al.*, 2012; Zhang *et al.*, 2014]. Such methods provide additional information about the face of a person, improving the accuracy of many facial classification tasks [Lu *et al.*, 2005; Uřičář *et al.*, 2012].

Drawbacks in prior methods for face analysis: Despite the success of previously proposed methods in inferring various personal traits from facial images, most solutions are based upon hand-designed features, and typically suffer from one or more of the following problems: (a) They are specifically tailored to a single task at hand [Su *et al.*, 2013; Tian and Bolle, 2003; Tjahyadi *et al.*, 2007; Hasan and Pal, 2014]; (b) They are not well scalable to real-world variations in data such as multiple view-points [Dhall *et al.*, 2011]; (c) They make use of unautomated pre-processing methods such as hand-labeling of key facial regions [Kumar *et al.*, 2009].

Major advantages of deep learning: Since deep learning based procedures can *automatically* learn a diverse set of low and high-level representations for the input data, they circumvent the need for building hand-crafted features. Also, since deep nets work directly on input images, there is seldom any need to do unautomated or esoteric preprocessing. Furthermore, it is well established [Krizhevsky *et al.*, 2012] that deep nets are robust to real-world variations in the input data, and thus are promising for building large-scale practical applications.

Applying Deep Learning to range of facial attributes: Given the promise of deep learning and the nature of our problem where we aim to predict attributes ranging from objective to subjective ones, from a diverse set of facial images, CNNs are an excellent fit for our needs. We thus apply CNNs for predicting the facial attributes. Previous papers on personal attribute prediction with deep nets have either not focused on facial attributes [Shankar *et al.*, 2015], or have only considered a very restricted set of facial attributes such as emotions [Lisetti and Rumelhart, 1998; Liu *et al.*, 2014]. Also, where researchers have tried to rank facial attributes for better classification [Parikh *et al.*, 2012; Shankar *et al.*, 2013], esoteric and plausibly subjective supervision is required.

Augmenting CNNs with face alignment information: While training, a CNN is inherently expected to learn features in a way which can correctly tell us about the spatial regions in the images most salient for the prediction of a class. For maximum robustness, these spatial regions should be consistent across all the training images of a given class. For instance, the personal traits exhibited in faces generally correspond to specific facial regions or a combination of them - hair color is mostly captured in the hair region of the face; happiness is specific to the region around lips; while old age can be seen as a combination of features around the forehead, under-eyes and cheeks. Thus, for all training images belonging to the class of hair-color, we would like that the CNN learns features that correspond to the hair regions of the image for prediction. If the CNN predicts white / blonde hair-color by considering the white skin color of a person, we would term that as erroneous. While a human can innately and consistently figure out such structural accordances in an image for a given class, the task is rather difficult for CNNs, more so when the classes are attributes (as against the objects). Noticing that the faces have a well-defined structure (forehead, eyes, noses, mouth, etc) which can be robustly captured using state-of-the-art techniques like [Zhang *et al.*, 2014], we augment the input data with this structural information to train a CNN. We thus expect it to learn more robust attribute-specific features, thereby ameliorating the prediction accuracy.[1]

1.1 Our Contribution:

We contribute a new Face Attributes Dataset (FAD), comprising of roughly 200,000 attribute labels for over 10,000 facial images. Our dataset covers many traits of individuals, and has labels regarding both objective and subjective personal attributes. The dataset has been carefully crowd-sourced from Amazon Mechanical Turk, establishing the veracity of the labels obtained.

We apply deep learning for predicting a wide range of facial attributes. We corroborate that using a CNN architecture for determining facial attributes provides an impressive baseline performance. To further improve performance, we propose an augmentation approach that incorporates facial landmark information for input images as an additional channel, helping the CNN learn better attribute-specific features so that the landmarks across various training images hold correspondence. We empirically show consistent improvement with our proposed approach over the aforementioned baseline across traits.

2 Face Attributes Dataset (FAD)

Our dataset consists of 10,000 facial images of celebrities (public figures), where each image is tagged with various traits of the individual. The images we used are a subset of the PubFig dataset [Kumar *et al.*, 2009].

The original PubFig dataset consisted of 60,000 images of celebrities, where each celebrity is covered by multiple images under different poses, at different times, and with a different expression. Due to copyright issues, original images were never provided for the PubFig Dataset, and only the respective internet addresses (URLs) were given. Since the release of PubFig, many of those URLs have become invalid, so we focused on the subset of images of the original data which are still available online.

The resolution of the 10,000 images downloaded was not constant. Since typically all the input images to a CNN are of the same size, we scaled each image to a fixed resolution of 150×150 pixels. We chose this resolution since most images posted on social media sites do not contain faces bigger than that (typically people pose with their torsos as well, if not the full body). Our dataset has thus been curated keeping practical applications in mind; so algorithms performing well on our dataset should also perform well on other real-world data.

2.1 Ground-Truth Annotations

As our target variables, we focused on multiple objective and subjective traits; the objective traits include: gender, ethnicity, age, make-up and hair color; the subjective traits include emotional expression, attractiveness, humorousness and chubbiness. The classes considered for each of these traits / attributes are listed in Table 1 along with their level of skewness. We emphasize that in this paper we consider the prediction of classes for each trait to be a discrete classification problem (e.g. we only aim to know whether a person's gender is male or female and not the degree to which they appear to be be masculine or feminine.)

[1] This can also be seen as a knowledge-transfer approach with deep learning (though in a different sense from transfer learning and multi-task learning methods employed with some deep nets [Zhou *et al.*, 2014; Oquab *et al.*, 2014; Zhang *et al.*, 2015].)

Trait	Data distribution
Gender	Male (50.8%), Female (49.2%)
Ethnicity	White (79.5%), Other (20.5%)
Hair Color	Dark (60%), Bright (40.0%)
Makeup	Wears (39.4%), Does not wear (60.6%)
Age	Young (67.8%), Elder (32.2%)
Emotions	Joy (64.2%), Other (35.8%)
Attractive	Yes (65.9%), No (34.1%)
Humorous	Yes (55.6%), No (44.4%)
Chubby	Yes (57.3%), No (42.7%)

Table 1: **Attributes / Traits in FAD:** Personal traits in FAD along with the corresponding classes are listed. For each trait, the distribution of images across the corresponding classes is given. As is evident, some traits have more skewness across their classes as compared to others. For all our experiments, our training and test sets contain a similar distribution.

In order to get the images labelled for various traits, we used Amazon's Mechanical Turk (MTurk). This is a crowdsourcing platform, which allows people to post micro-tasks, and lets participants fulfill these tasks for a fee. We sourced a total of 1,500 raters from MTurk. All the participants were sourced from the US and Canada. We let each of the participants examine several images and provide labels for each image for each of the traits listed in Table 1.

We offered each MTurk participant a payment of $6 for filling in all the trait labels for 10 of our images. To account for the fact that all participants on MTurk might not exert enough effort (or be satisfactorily sincere) in the annotation task, we made sure we have enough non-redundant labels, by having each image labelled 3 times.

To further ensure the quality of the labels, we included some very simple questions designed to identify participants who could be randomly clicking answers or not paying enough attention to the task.[2] We removed the responses of participants who failed to correctly answer these questions. Also, we excluded the responses of participants who disagreed with their peers on over a third of the labels for the objective traits (e.g. participants who did not agree with their peers on the gender or ethnicity labels for a third of their images).

Our goal is to use the annotations of the images in FAD to train an automated system to infer personal traits from facial images. However, some traits are clearly more difficult than others. When people find it easy to infer a certain property from an image, we expect a high degree of agreement between the raters. In contrast, when inferring a target variable is difficult, we expect our annotators to often disagree regarding the correct label for an image.

In cases where a trait exhibits a low degree of inter-rater agreement in the dataset, even an excellent learning method would find it difficult to achieve a high degree of accuracy in the task. Table 2 presents the inter-rater agreement, as measured by Fleiss' Kappa, for each of the traits, evaluated on our dataset.

Trait	Data distribution
Gender	0.9601 (APA)
Ethnicity	0.913 (APA)
Hair Color	0.719 (SA)
Makeup	0.697 (SA)
Age	0.563 (MA)
Emotions	0.688 (SA)
Attractive	0.29 (FA)
Humorous	0.171 (SLA)
Chubby	0.153 (SLA)

Table 2: **Inter-rater agreement (Fleiss' Kappa) measured for each of the traits in FAD.** A value of 1 indicates perfect agreement, while a value of 0 indicates no agreement. APA stands for *Almost perfect agreement*; MA for *Moderate agreement*; SA for *Substantial agreement*; FA for *Fair agreement*; and SLA for *Slight agreement*.

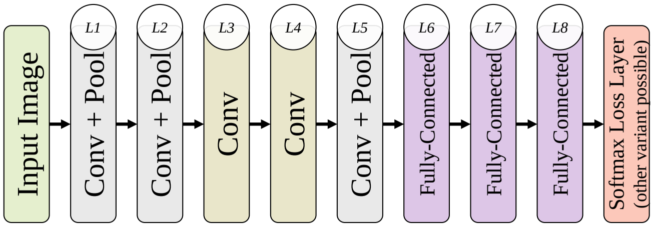


Figure 1: **Block Illustration of AlexNet [Krizhevsky *et al.*, 2012]:** a deep CNN architecture. The deep convolutional neural net has eight layers (denoted as $L1, \ldots, L8$) after the input. The last fully connected layer is conventionally followed by a softmax loss layer, but can also be replaced by the likes of Sigmoid Cross Entropy Loss Layer [Jia *et al.*, 2014]. We use this CNN architecture for all our experiments.

3 Prediction Algorithm

Our goal is to predict the various traits of a person from their facial images. We consider FAD for all our experiments, and use the traits and the corresponding classes as listed in Table 1. Due to the various advantages offered by deep nets, we begin by applying one of the most famous CNN architectures [Krizhevsky *et al.*, 2012] for our prediction task, widely known as AlexNet. The block-level architecture of AlexNet is shown in Figure 1.

Brief Overview of AlexNet: The fully-connected layers have 4096 neurons each. Max-pooling is done to facilitate local translation-invariance. For the fully connected layers, a drop-out [Srivastava *et al.*, 2014] probability of 0.5 is used to avoid overfitting. The final fully connected layer takes the outputs of $L7$ as its input, produces outputs equal to the number of classes through a fully connected architecture, then passes these outputs through a softmax function, and finally applies the negative log likelihood loss. With softmax loss layer, each input image is expected to have only

[2] For example, we asked simple mathematical questions for which every participant is expected to know the answer, such as "how much is 6+8?"

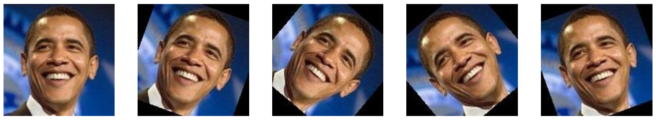


Figure 2: **Rotation of images in the training set:** In order to make the deep net training more robust to facial pose variation, for each training image (left most), we create 4 new training images as its rotated versions (last four). The original training image is rotated by $\{-40, -20, 20, 40\}$ degrees.

one label. When the softmax loss layer is replaced by a sigmoid cross-entropy loss layer, the outputs of $L8$ are applied to a sigmoid function to produce predicted probabilities, using which a cross-entropy loss is computed. Here each input can have multiple label probabilities. We refer the reader to [Krizhevsky *et al.*, 2012] for complete details of AlexNet.

Choice of the Loss Function: Softmax Loss and Sigmoid Cross-Entropy Loss are the two most widely used loss functions for classification tasks in deep learning. With the softmax loss layer, the training of the AlexNet is typically accomplished by minimizing the following cost or error function (negative log-likelihood):

$$\mathcal{L}_s = -\frac{1}{N}\sum_{r=1}^{N}\log(\hat{p}_{r,y_r}) + \mathcal{L}_R \tag{1}$$

where r indexes N training images across all traits ($r \in \{1,\ldots,N\}$), $\mathcal{L}_R = \lambda||\boldsymbol{W}||_2$ is the L2 regularization on weights $\boldsymbol{W}$ of the deep net, λ is a regularization parameter, and the probability $\hat{p}_{r,y_r}$ is obtained by applying the softmax function to the M outputs of layer $L8$, M being the number of classes we wish to predict labels for. Letting $l_{r,m}$ denote the m^{th} output for r^{th} image, we have

$$\hat{p}_{r,m} = \frac{\mathrm{e}^{l_{r,m}}}{\sum_{m'}\mathrm{e}^{l_{r,m'}}}, \qquad m, m' \in \{1,\ldots,M\}. \tag{2}$$

In case one applies the sigmoid cross entropy loss, each image is expected to be annotated with a vector of ground-truth label probabilities $\boldsymbol{p}_r$, having length M, and the network is trained by minimizing the following loss objective:

$$\mathcal{L}_e = -\frac{1}{NM}\sum_{r=1}^{N}\sum_{m=1}^{M}[\boldsymbol{p_r}\log(\boldsymbol{\hat{p}_r}) + (1-\boldsymbol{p_r})\log(1-\boldsymbol{\hat{p}_r})] + \mathcal{L}_R \tag{3}$$

where the probability vector $\boldsymbol{\hat{p}_r}$ is obtained by applying the sigmoid function to each of the M outputs of layer $L8$.

A natural choice to approach our prediction task is to train a single CNN for all our traits / attributes. Since an image can have multiple traits, the sigmoid cross-entropy loss function is best suited for our scenario. However, we find that for our prediction task where for every given trait, we have mutually exclusive attribute classes, training one net for each given trait provides a higher accuracy. We thus establish our baseline and perform all our experiments with the latter choice. For a greater number of facial traits, one can combine the features from these independently trained CNNs, to train some fully connected layers.

Rotating images in the training set: In order to make the deep net training more robust to facial pose variation, for each training image, we create 4 new training images as its rotated versions. Each training image is rotated by $\{-40, -20, 20, 40\}$ degrees. This also increases the size of our training set by a factor of 5. Example of the rotated versions of an input image is shown in Figure 2.

3.1 Incorporating Facial Landmark Information

For each of the traits, we train the network on FAD using labels for the classes of that trait, and evaluate its performance as a baseline. The input to the network is a color image, in a resolution of $D_x \times D_y$ (we used $D_x = D_y = 150$ pixels); each pixel is represented as a three "channel" RGB encoding. Thus the input layer has 3 neurons, each neuron representing a 2-D matrix of size $D_x \times D_y$.

Our improved approach is based on incorporating facial landmark information in the input data. Localizing facial landmarks, sometimes referred to as "face alignment" is a key step in many facial image analysis approaches.

Various recognition algorithms, including those dealing with facial figures, require exact positioning of an object into a canonical pose, to allow examining the position of features relative to a fixed coordinate system. Inspired by such methods, we embed landmark information in deep nets for predicting a wide range of facial attributes.

Facial landmark localization algorithms are designed to find the location of several key "landmarks" in an image, such as the location of the center of the eyes, parts of the nose or the sides of the mouth. Consider a list $L = (l_1, \ldots, l_k)$ of facial landmarks. Facial landmark localization algorithms receive a facial image I as an input, and output the coordinates in the image for each of the landmarks $C^I = (c_1^I, \ldots, c_k^I)$ where $c_j^I = (x_j^I, y_j^I)$ are the coordinates of landmark l_j in the image I. An example of an image and the corresponding facial landmarks is given in Figure 4.

Our approach uses a facial landmark localization algorithm as a subroutine, so any such algorithm could be used by our approach. It operates by associating each pixel in the facial image with the *closest* facial landmark for that image. We then add this association as an additional channel to each input image.

We now formally describe our approach. In our baseline approach, the pixel in coordinate (x, y) in the input image I is encoded as three RGB channels $(R^{(x,y)}, G^{(x,y)}, B^{(x,y)})$. We add an additional channel, relating to the closest facial landmark, denoted as $A^{(x,y)}$, thus increasing the number of neurons in the input layer from 3 to 4. $A^{(x,y)}$ encodes the identity of the nearest facial landmark to the pixel in coordinates (x, y).

To compute $A^{(x,y)}$, we call the facial landmark localization algorithm (FLL) as a subroutine, to obtain a list of landmark coordinates, $C^I = (c_1^I, \ldots, c_k^I)$ where $c_j^I = (x_j^I, y_j^I)$, and compute the distance between pixel (x, y) to each of these coordinates, to obtain $d_j^I(x, y) = ||(x_j^I, y_j^I), (x, y)||_2$.

We select the index of the facial landmark nearest to the pixel as the value of the pixel in the additional channel. Finally, we train the CNN on the set of augmented images, consisting of the original RGB channels and the new channel encoding the nearest landmark associated with each

```
procedure AUGMENT-FLL
    $(I = (R^{(x,y)}, G^{(x,y)}, B^{(x,y)}))$
    $(c_1^I, \ldots, c_k^I) = FLL(I)$ // Get facial landmarks
    for $x = 1$ to $D_x$ do
        for $y = 1$ to $D_y$ do
            for $j = 1$ to $k$ do
                $d_j^I(x, y) = ||(x_j^I, y_j^I), (x, y)||_2$
                // pixel-landmark distances
            $A^{(x,y)} = \arg\min_{j \in \{1,\ldots,k\}} d_j^I(x, y)$
    return $I' = (R^{(x,y)}, G^{(x,y)}, B^{(x,y)}, A^{(x,y)})$
```

Figure 3: **Creating $A^{(x,y)}$:** Algorithm for encoding the identity of the nearest facial landmark to every pixel. This algorithm is used to create the additional channel $A^{(x,y)}$ which augments the input images.

Trait	Baseline	LACNN
Gender	**98.46%**	98.33%
Ethnicity	82.7%	**83.35%**
Hair Color	91%	**91.69%**
Makeup	92.5%	**92.87%**
Age	88.42%	**88.83%**
Emotions	**88.93%**	88.33%
Attractive	78.44%	**78.85%**
Humorous	66.8%	**69.06%**
Chubby	60.6%	**61.38%**

Table 3: **Comparison of prediction accuracy:** The accuracy of the baseline CNN method and LACNN on FAD. For most of the objective and subjective traits, LACNN improves the prediction accuracy.

pixel. We refer to our approach as the **L**andmark **A**ugmented **C**onvolutional **N**eural **N**etwork (LACNN) method.

The algorithm for generating the additional channel is given in Figure 3. For facial landmark detection, we have used the state-of-the-art TCDCN face alignment tool [Zhang *et al.*, 2014], which returns the locations of $k = 68$ key facial landmarks. In AUGMENT-FLL, TCDCN is thus used for FLL. We find that TCDCN is fairly robust to facial viewpoint variation. Instead of TCDCN, any other facial landmark detection tool could also be used. An illustration of the augmented input channel $A^{(x,y)}$ is shown in Figure 4.

4 Results

We now discuss the performance of the baseline CNN approach and LACNN. As mentioned before, we use FAD for all our experiments.

For training and inference with CNNs, we have used the Caffe Library [Jia *et al.*, 2014]. For doing inference on a considerable amount of test images, we create a 80/20 train/test split with FAD, maintaining the same data distribution across the training and test sets for all traits as given in Table 1. Such a split evaluates our method on roughly 36,000 labels.

Table 3 shows the accuracy of the baseline CNN method and LACNN on FAD. It is clear that CNN provides an overall

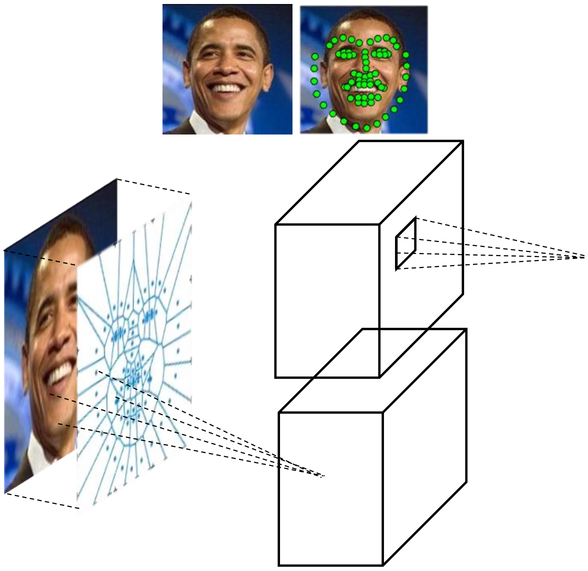


Figure 4: **Using facial landmarks for an input image: (*Top*)** Example of an input image and the corresponding facial landmarks detected using TCDCN [Zhang *et al.*, 2014]. **(*Bottom*)** An illustration of the augmented input channel $A^{(x,y)}$ computed using AUGMENT-FLL. Each region in $A^{(x,y)}$ is coded with a different value as per the index of the associated landmark. The figure shows the additional input channel being fed into the subsequent parts of convolutional neural network, along with the RGB image.

impressive baseline performance. Even for highly subjective traits, where human raters tend to disagree regarding the correct label of an image (see Table 2), CNNs give a reasonable performance. This indicates that CNN based approaches are indeed flexible, and can handle many traits without resorting to building ad-hoc systems relying on hand-crafted features.

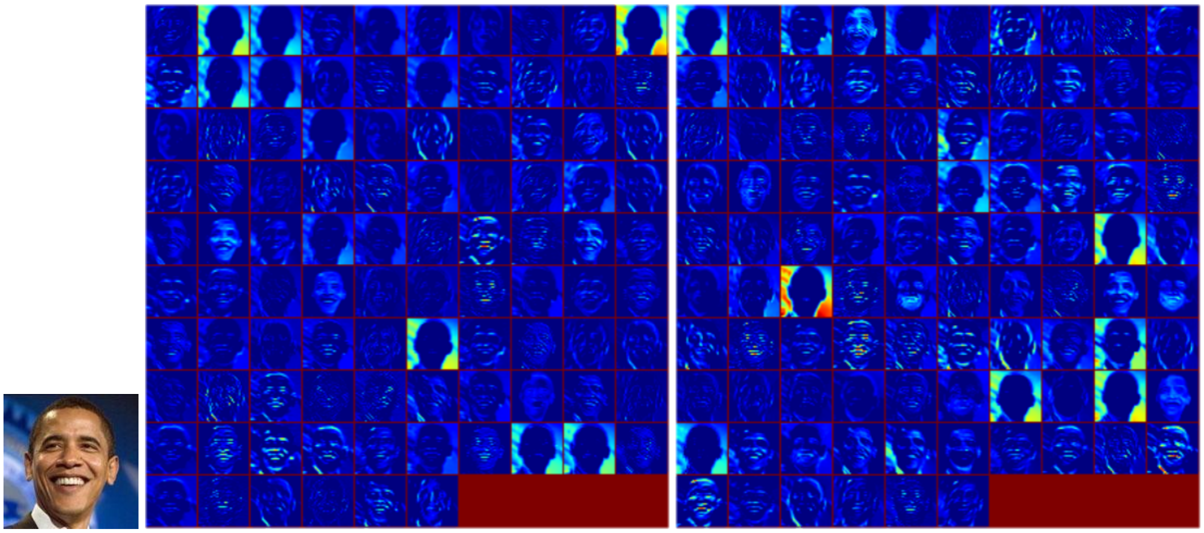
The proposed LACNN shows consistent improvement across most of the traits as compared to the CNN baseline. Note that LACNN has the capability to improve performance for both the objective as well as the subjective traits. This substantiates our intuition that face alignment information can be useful in predicting facial attributes using deep nets.

To further validate our intuition that facial landmarks should help the CNN learn more robust attribute-specific features in a more consustent manner, we depict the visualizations of the output responses of the first convolutional layer of AlexNet, trained with both baseline CNN and LACNN in Figure 5.

Observing the filter activations of the first convolutional layer shows that the responses generated using LACNN have more detailed information as compared to the ones generated with the baseline of a non-augmented CNN.

The output responses generated with LACNN have many variations of prominent facial parts including the nose, eyes, hairline, etc. Further, there is a higher number of neurons exhibiting such valuable information in the case of LACNN.

The outputs with better discernible information can be attributed to the fact that landmark augmentation helps the



(a) Input (b) Responses using baseline CNN (c) Responses using LACNN

Figure 5: **Visualizing output of the first convolutional layer of AlexNet trained with baseline CNN and LACNN:** For an input image, figure shows the outputs responses of the first convolutional layer of AlexNet trained using the baseline CNN method and our proposed LACNN. Both visualizations have been generated from nets trained on the same trait. The first convolutional layer of AlexNet contains 96 neurons, whose outputs are shown here on a 10 × 10 grid. A careful observation reveals that the responses generated using LACNN contain more detailed information as compared to the ones generated with baseline CNN: neurons in LACNN contain more discernible information about key facial parts. Also, more neurons exhibit valuable information in LACNN.

CNN to learn filters in a way such that similar regions across a range of facial images hold correspondence to exhibit similar responses. This is clearly important for facial attribute prediction, since a given trait in any face is always associated with the same combination of facial sub-parts.

5 Conclusions and Future Work

We have proposed a method for predicting personal attributes from facial images, based on a CNN architecture augmented with face alignment information. We have empirically evaluated our approach by building a tagged facial images dataset called FAD, showing that improved classification performance can be achieved for a very wide range of traits using our approach.

Several questions remain open for further research. First, could one devise a method using facial landmarks to better detect facial attributes, such that attribute-specific regions are explicitly learned for faces? Could such a method be used for *ranking* facial images according to attributes?

Further, could one detect more subjective attributes such as more detailed emotions or traits such as being in shape (muscle tone) or other health related traits or friendliness? Could such an analysis be based on the information contained in the nets trained for the basic objective attributes?

Finally, could one exploit graph-structured compositions within the deep nets to better interpret facial traits? More generally, a key disadvantage of CNN based methods is that the learned model is not “human interpretable” in the sense that it is difficult to understand which sub-parts of the network drive the prediction. Would it be possible to train multiple nets or a single net for many traits and examine the correlation, so that it would be possible to explain the predictions made by the system in a way understandable by humans?

References

[Behnke, 2003] Sven Behnke. *Hierarchical neural networks for image interpretation*. Science & Business Media, 2003.

[Bengio *et al.*, 2007] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. *NIPS*, 2007.

[Datta *et al.*, 2008] Ritendra Datta, Jia Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, pages 105–108. IEEE, 2008.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[Dhall *et al.*, 2011] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *FG*, 2011.

[Fasel and Luettin, 2003] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[Hasan and Pal, 2014] Md Kamrul Hasan and Christopher Pal. Experiments on visual information extraction with the faces of wikipedia. In *AAAI*, 2014.

[Horng *et al.*, 2001] Wen-Bing Horng, Cheng-Ping Lee, and Chun-Wen Chen. Classification of age groups based on facial features. *TJSE*, 2001.

[Hosoi *et al.*, 2004] Satoshi Hosoi, Erina Takikawa, and Masato Kawade. Ethnicity estimation with facial images. In *AFGR*, 2004.

[Huang *et al.*, 2007] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678. ACM, 2014.

[Kagian *et al.*, 2006] Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-Or, and Eytan Ruppin. A humanlike predictor of facial attractiveness. In *NIPS*, pages 649–656, 2006.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Kumar *et al.*, 2009] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*. IEEE, Oct 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lisetti and Rumelhart, 1998] Christine L Lisetti and David E Rumelhart. Facial expression recognition using a neural network. In *FLAIRS*, 1998.

[Liu *et al.*, 2014] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812. IEEE, 2014.

[Lu and Jain, 2004] Xiaoguang Lu and Anil K Jain. Ethnicity identification from face images. In *Defense and Security*, 2004.

[Lu *et al.*, 2005] Xiaoguang Lu, Hong Chen, and Anil K Jain. Multimodal facial gender and ethnicity identification. In *Adv. Biometrics*. 2005.

[Lyons *et al.*, 1999] Michael J Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *TPAMI*, 1999.

[Maini and Aggarwal, 2009] Raman Maini and Himanshu Aggarwal. Study and comparison of various image edge detection techniques. *IJIP*, 2009.

[Microsoft, 2015] Microsoft. Guess my Age. `https://www.microsoft.com/en-us/store/apps/guess-my-age/9nblggh3t1ld`, 2015.

[Moghaddam and Yang, 2002] Baback Moghaddam and Ming-Husan Yang. Learning gender with support faces. *TPAMI*, 24(5):707–711, 2002.

[Oquab *et al.*, 2014] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724. IEEE, 2014.

[Padgett and Cottrell, 1997] Curtis Padgett and Garrison W Cottrell. Representing face images for emotion classification. *NIPS*, pages 894–900, 1997.

[Parikh *et al.*, 2012] Devi Parikh, Adriana Kovashka, Amar Parkash, and Kristen Grauman. Relative attributes for enhanced human-machine communication. In *AAAI*, 2012.

[Segundo *et al.*, 2010] Maurício Pamplona Segundo, Luciano Silva, Olga Regina Pereira Bellon, and ChauãC Queirolo. Automatic face segmentation and facial landmark detection in range images. *Systems, Man, and Cybernetics*, 2010.

[Shankar *et al.*, 2013] Sukrit Shankar, Joan Lasenby, and Roberto Cipolla. Semantic transform: Weakly supervised semantic inference for relating visual attributes. In *ICCV*, 2013.

[Shankar *et al.*, 2015] Sukrit Shankar, Vikas K. Garg, and Roberto Cipolla. Deep-carving: Discovering visual attributes by carving deep neural nets. In *CVPR*, June 2015.

[Simard *et al.*, 2003] Patrice Y Simard, Dave Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, page 958. IEEE, 2003.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[Su *et al.*, 2013] Roger Su, Timothy Michael Dockins, and Manfred Huber. Ica analysis of face color for health applications. In *FLAIRS*, 2013.

[Szegedy *et al.*, 2014] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. 2014.

[Tian and Bolle, 2003] Ying-li Tian and Ruud M Bolle. Automatic detecting neutral face for face authentication and facial expression analysis. In *Intelligent Multimedia Knowledge Management*, 2003.

[Tjahyadi *et al.*, 2007] Ronny Tjahyadi, Wanquan Liu, Senjian An, and Svetha Venkatesh. Face recognition via the overlapping energy histogram. In *IJCAI*, 2007.

[Uřičář *et al.*, 2012] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of facial landmarks learned by the structured output svm. *VISAPP*, 2012.

[Vinciarelli *et al.*, 2009] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *IVC*, 2009.

[Zhang *et al.*, 2014] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multitask learning. In *ECCV*. 2014.

[Zhang *et al.*, 2015] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 2015.

[Zhou *et al.*, 2014] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, 2014.

[Zhu and Ramanan, 2012] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.