Deep Cross-modal Age Estimation

Ali Aminian and Guevara Noubir

Northeastern University, Boston MA 02115, USA,
aliiaminian@ccs.neu.edu,
noubir@ccs.neu.edu

Abstract. Automatic age and gender classification systems can play a vital role in a number of applications including a variety of recommendation systems, face recognition across age progression, and security applications. Current age and gender classifiers, are lacking crucial accuracy and reliability in order to be used in real world applications since most real-time systems have zero fault tolerant. This paper develops an end-to-end, deep architecture aiming to improve the accuracy and reliability of the age estimation task.

We designed and trained a deep convolutional neural network (CNN) architecture for age estimation that builds upon a gender classification model. The system leverages a gender classifier to improve the accuracy of the age estimator. We investigate several architectures and techniques for the age estimator model with cross-modal learning, including an end-to-end model, using gender embedding of the input image, which leads to an increased accuracy. We evaluated our system on the Adience benchmark, which consists of real-world in-the-wild pictures of faces. We have shown that our system outperforms state-of-the-art age classifiers, such as [1] by 9%, by training a cross-modal age classifier.

Keywords: age and gender classification, convolutional neural networks

1 Introduction

Age and gender classification plays an important role in numerous applications, from a variety of computer vision-based recommendation and human-computer interaction systems. In these systems, age is considered an essential factor, from improving face recognition across age progression, to understanding relations between individuals, and making further inferences towards grasping social interactions. Security systems can also benefit from a better understanding of the age, gender, and capabilities of actors.

Recent developments in deep learning enabled the improvement of a variety of tasks. For instance, in computer vision, it is possible to achieve very high accuracy for applications, such as face detection, and face recognition [2]. This progress benefited from the confluence of new techniques in deep neural net (DNN) architectures, the presence of highly parallel low-cost computing infrastructure, such as GPGPU, and the availability of feature rich large datasets.

Despite this phenomenal progress, the performance of age classifiers is still lacking reliability and accuracy [1]. Age classification is an intrinsically challenging problem, as it is hard for humans to achieve with high accuracy. Furthermore, it is difficult to obtain massive accurately-labeled datasets for pictures in-the-wild, unlike other computer vision tasks, such as object classification.

In this paper, we propose and investigate an approach to improve the accuracy of age classification with cross-modal learning by using a pre-trained gender classifier. In addition to training the age classifier with the same deep model as we have used for gender classification, we also feed the model with the gender embedding as the gender classifier extracts useful features from the gender. This approach is motivated by the fact that the aging process is not identical in men versus women, due to biological and behavioral differences. For instance, skin thikness, collagen density, rate of collagen loss at different stages of life, texture, and level of tissue hydration [3].

Beyond a straightforward integration of the gender classifier output, we explored several techniques to improve the system accuracy, by feeding the embedding of an improved model of each task to the other task. The aim being that the deep model is able to find more essential features at each round. We considered choosing the gender embedding from different fully connected (FC) layers by integrating it into different layers of the age classification model. We also considered an end-to-end model of a more general integration of the two classifier networks. We evaluated our model on the Adience dataset, a thorough benchmark for age and gender classification [4]. The Adience dataset consists of in-the-wild unfiltered face images, that present the typical variations in appearance, noise, pose, and lighting expected of images taken without careful preparation or posing. We selected the Adience dataset because of its realism, although it creates additional challenges. Our evaluation results demonstrate that our techniques outperform state-of-the-art approaches in age classification.

To summarize our contributions, we designed a simple architecture, used as the underlying model for both age and gender classification. We designed, trained, and evaluated several cross-modal learning models for age classification built upon gender classification model. In addition, we proposed an end-to-end model which automates the cross-modal learning process during the training. We also iteratively refined both models, in order to improve accuracies of both tasks.

2 Related works

Before describing our approach, detailed design, and performance, we briefly summarize the related work, both for DNNs, as well as age and gender classification.

2.1 Age and Gender Classification

Gender Classification. The problem of gender classification from facial images received significant attention in recent years and many approaches have been developed for this purpose. A survey of gender classification can be found in [5] and more recently in [6]. Below, we briefly survey relevant methods.

Early methods for gender classification used a neural network on a small set of near-frontal face images [7]. Later work used the 3D structure of the head for classifying gender [8]. SVM classifiers were investigated in [9]. Other works used AdaBoost for gender classification [10]. Finally, viewpoint-invariant age and gender classification were introduced in [11].

More recently, Webers Local texture Descriptor [12] were used in [13] for gender recognition on The Face Recognition Technology (FERET) benchmark [14]. In [15], intensity, shape and texture features were used again on the FERET benchmark.

FERET benchmark [14] developed by the DoD to facilitate face recognition has been a popular performance evaluation method. It worth noting that the FERET dataset was developed under highly controlled conditions. Hence, FERET images are less challenging than in-the-wild face images. Furthermore, due to its extensive use for gender classification evaluation, the FERET benchmark is saturated. Therefore, realistic comparison of these techniques became a difficult task. Recent work started evaluating using newer datasets. In [16], a combination of LBP with an Adaboost classifier approach, experimented on the popular Labeled Faces in the Wild (LFW) [17] benchmark. However, the main usage of LFW is face recognition.

Due to recent advances in deep models, several methods have been proposed with significant improvement for age prediction. For instance, [18–20] developed deep models for classification. [21] proposed a network for both gender and smile prediction.

Age Classification. The problem of age classification from facial attributes, also recently attracted significant attention, due to its usefulness in real-world applications. A detailed survey of age classification can be found in [22], and [23], and more recently in [24].

Early methods used extraction of ratios [25], given the facial landmarks for each image. To this end, [26] used a similar method for age progression. Since all these methods need an accurate localization, the benchmark being used is highly controlled for constrained images.

In another line of work, few methods represent the aging process as a subspace [27] or a manifold [28]. Since these methods require to have a near-frontal faces, therefore, these methods are developed on constrained datasets with near-frontal faces (e.g UIUC-IFP-Y [29, 28], FG-NET [30], and MORPH [31]). Similar to early approaches, these methods are also inadequate for in-the-wild datasets.

Aside from the approaches described above, some methods used local features for representing face images. In [32], Gaussian Mixture Models (GMM) [33] were used for representing the distribution of local patches. In [34], GMM were used, but instead of pixel patches, robust descriptors were used. On the other hand, instead of GMM, Hidden-Markov-Model, super-vectors [35], were used in [36] for representing face patch distributions.

As an alternative to previous methods, [37] used Gabor image descriptors [38]. In [39], a combination of Biologically-Inspired Features (BIF) [40] and various manifold-learning methods were used for age estimation. The Gabor image descriptor [38], and local binary patterns (LBP) [41] were used in [42] with a hierarchical age classifier based on SVM [43].

Moreover, [44] proposed the improved versions of relevant component analysis [45] and locally preserving projections [46], which are used for dimensionality reduction with Active Appearance Models (AAP) [47] as an image feature. Up to this point, the best performing methods were demonstrated on the Group Photos benchmark [48]. All of these methods have proven effective on small and/or constrained benchmarks for age estimation.

Aside from datasets that are taken under highly controlled condition, AgeBD [49], is the first manually collected in-the-wild age database. However, in this paper, similar to [1], we focus on the more challenging, in-the-wild Adience benchmark, for instance in comparison to LFW [17]. We train and report our system performance on this challenging dataset as well.

Finally, researchers leveraged deep neural networks to achieve better results on facial age classification tasks. In [50], a deep model for real and apparent age prediction was proposed using the VGG-16 architecture, and was trained on ImageNet [51]. In addition, [52–55] achieved significant improvement in age prediction by using deep architectures. The work by [1] leveraged recent developments in DNNs, and outperformed all previous methods by training a simple, and independent deep network for both age and gender classification on Adience dataset. However, these methods are still far behind the human accuracy, due to their intrinsic simple architecture, and the fact that age and gender are trained separately. In this work, we show that our proposed method which leverages deep CNN, with a cross-modal learning approach, outperforms all previous methods for age classification.

2.2 Deep Neural Networks

A detailed survey of recent advances in CNNs can be found in [56]. LeNet-5 network described by [57] for optical character recognition is considered as one of the first applications of CNNs. However, the network was relatively modest due to the limited computational resources in comparison with more modern deep networks. AlexNet was introduced [58] and significantly improved the classification accuracy in the ImageNet competition.

To name a few applications, human pose estimation [59], face parsing [60], facial key-point detection [61–63], face analysis [64, 65], speech recognition [66], and action classification [67], all benefited from recent advances in the design of deep CNNs.

More recently, beyond single deep architecture methods, the emergence of cross-modal deep architectures [68], and [24], resulted in improvements where models for one task can built upon another task model.

3 Background

In this section, we first present how age and gender are associated, and can provide useful information for each other in classification task, and then we discuss our high-level designs, and the underlying architecture.

3.1 Age and Gender Association

In our context, age and gender classification are associated in the sense that the output of each potentially contain valuable information for the other task. To illustrate, the skin on a man versus that on a woman is significantly different. Aside from the ability to grow a beard, which is one obvious example, from a structural point of view, some of the differences include skin thickness, collagen density, loss of collagen as we age, texture and hydration. Thickness of the skin varies with the location, age and sex of the individual. In addition, androgens (i.e. testosterone), which cause an increase in skin thickness, accounts for why a man's skin is about 25 percent thicker than that of a woman's [3]. Consequently, knowing that an individual is male in advance, can help the network better identify age category due to specific attributes that are common in specific gender at each age.

3.2 Age Classification Approach

Our proposed architecture is based on the work by [1], which a simple CNN architecture was used to independently classify age and gender, given the input image. The model consists of three convolutional layers and two FC layers. Despite its simplicity, it improved the state-of-the-art age and gender classifiers on the challenging newly released Adience dataset by 5%.

We propose to extend this prior work by cross-modal learning, training an age estimation model based upon a pre-trained gender classifier. We feed the extracted embedding of the pre-trained gender classifier to our age classification model during the training phase. In other words, we help our model to extract better features for age classification during the training phase, given the gender embedding from a pre-trained model. In addition, the opposite is true, since knowing the age can provide helpful information to detect the gender. We considered several variants of the embedding and feeding in our system, and finally proposed a deep end-to-end model which automatically chooses the best variant.

In this work, first we re-train a gender classifier following [1] with the same architecture. Afterwards, we train our age classification model, given the image and gender embedding from the pre-trained gender classifier. Finally, we test our model following the same steps in [1] and other works, to compare our results and accuracy with proposed systems under the same conditions. We explain the detailed architecture, as well as our proposed designs in the next section. Figure 1 illustrate the cross-modal learning idea in CNNs.

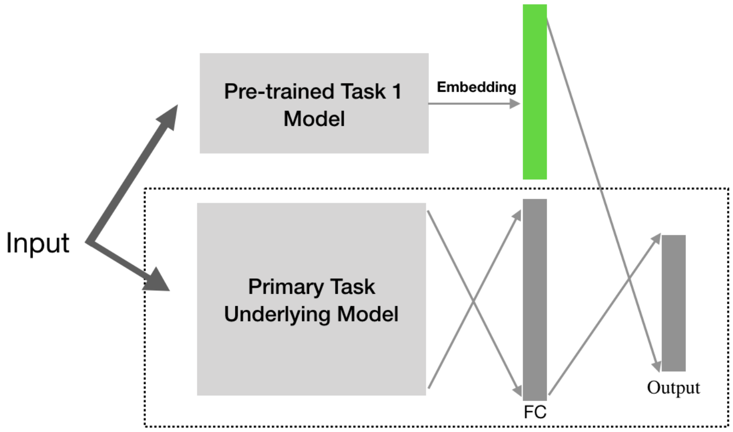

Fig. 1. Cross-modal learning. This shows how the cross-modal learning works in general. We consider the model as a black box, with any arbitrary design. We have a pre-trained model for task one. After obtaining the embedding of the input from pre-trained model, we can concatenate the embedding at some layer in our primary model, aiming to obtain better results in the training task

4 Technical Approach

In this section, we first present our model architecture along with possible designs and architectures, discuss the training and testing process, and then the iterative model refinement technique and its improvement in regard to accuracy.

It worth mentioning that since access to personal information on the subjects in the images is limited, we have available datasets limited in size compared to face and object detection datasets. Therefore, we are exposed to the risk of over-fitting. As a consequence, we avoid over-fitting by having fewer layers and neurons, as well as having dropout layers.

4.1 Model Architecture

We use the CNN network architecture introduced in [1] as the basis for our work. This underlying architecture is illustrated in Figure 2 and it follows the conventions in AlexNet [58] architecture. It consists of three convolutional layers, followed by two FC layers, with a small number of neurons. By comparing this simple network to other architectures, it is clear that the proposed network has fewer layers and neurons. The choice of a network with a smaller number of neurons and layers is motivated by the desire to avoid over-fitting, as well as the nature of the problem. That is because the output is of size eight, which is fairly small compared to other classification problems such as object detection, which consists of tens of thousands of output categories. Since we follow the AlexNet architecture conventions, we use the same parameters and hyper parameters for the first three convolutional layers. For the last two FC layers, we choose 512 as the number of neurons, in order to avoid over-fitting.

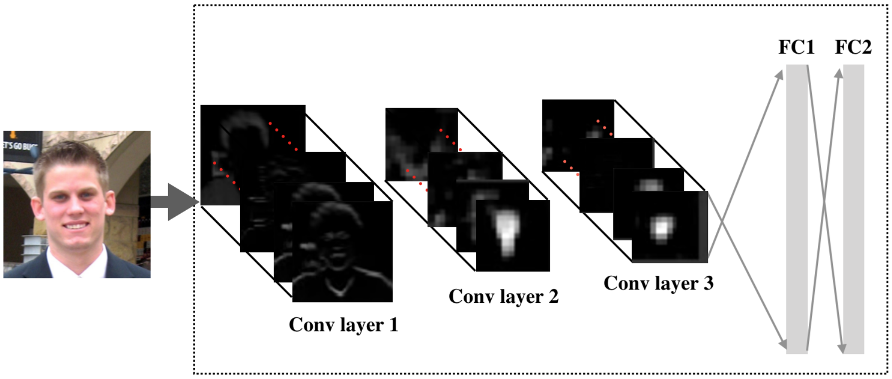

Fig. 2. Illustration of our underlying CNN architecture. The model contains three convolutional layers, each followed by a rectified linear operation and pooling layer. The first two layers also follow normalization using local response normalization [58]. The first convolutional layer contains 96 filters of 7×7 pixels, the second convolutional layer contains 256 filters of 5×5 pixels, The third and final convolutional layer contains 384 filters of 3×3 pixels. Finally, two FC layers are added, each containing 512 neurons

All three color channels are considered and processed through the network. Each input image is first re-scaled to 256×256, and then a crop of 227×227 is fed to the network depends on the proper cropping method. Below are the details of three convolutional layers.

Layer 1: 96 filters of size $3 \times 7 \times 7$ pixels applied to the input image, and then followed by a rectified linear operator (ReLU). Then, a max pooling layer, which takes maximum value of each 3×3 region, with 2 pixels strides (this is the distance between the receptive field centers of neighboring neurons in a kernel map), followed by a local response normalization layer. On the first convolutional layer, we used neurons with receptive field size $F = 11$, stride $S = 4$ and no zero padding $P = 0$. Since $(227 - 11)/4 + 1 = 55$, and because we have 96 filters ($K = 96$), the convolutional layer output volume has the size $96 \times 55 \times 55$. Each of the $96 \times 55 \times 55$ neurons in this volume was connected to a region of size $3 \times 11 \times 11$ in the input volume. Furthermore, all 96 neurons in each depth column are connected to the same $3 \times 11 \times 11$ region of the input, but with different weights. Since max pooling layer region is 3×3, with $S = 2$, then we have $(55 + 1)/2 = 28$ for both W and H. Therefore, we have an output of size $96 \times 28 \times 28$ due to 96 filters that we applied.

Layer 2: From the previous convolutional layer, we have $96 \times 28 \times 28$ output, which would be considered as the input of this layer. This layer contains 256 filters of size $96 \times 5 \times 5$ pixels, with stride one ($S = 1$), and the same padding, which is followed by ReLU, max pooling layer, and then again local response normalization, with the same hyper parameters as before. Hence, we have $256 \times 28 \times 28$ as the output of convolutional layer, and since $S = 2$, we have $256 \times 14 \times 14$ as this step's output.

Layer 3: Finally, last convolutional layer operates on a $256 \times 14 \times 14$ blob. This layer contains 384 filters of size $256 \times 3 \times 3$ pixels (3×3 filters), followed by ReLU and a max pooling layer, again, with the same hyper parameters as before. It worth noting that, again, we have stride of size one ($S = 1$), and the same padding as the previous convolutional layer. As a consequence, the output is of size $384 \times 6 \times 6$.

Then, the FC layers are defined as below.

Layer 4: This layer receives the output of the third convolutional layer ($384 \times 6 \times 6$), and contains 512 neurons, which is followed by a ReLU and dropout layer. We point out that the output of the last convolutional layer would be reshaped to a single array of size 13824.

Layer 5: This layer receives the output of first FC layer which has 512 dimensions, and similarly contains 512 neurons followed by a ReLU and dropout layer.

4.2 Prediction Head and Cross-modal learning

Potentially, we can feed a softmax classifier by the output of the last FC layer, and consider each number as the probability of the corresponding category. However, in this work, for our prediction head, given the input image, we use the last FC layer and the gender embedding from the pre-trained gender classifier.

There are several variants for choosing gender embedding since the last few layers in the gender classification model is encoding the image at different levels. Intuitively, the first FC layer has extracted less abstract features; however, the second FC layer contains more complex features embedded. In this design, we use three separate architectures, which we discuss below.

FC Layer Embedding of Gender into Age Classifier (Manual Embedding). The age classifier network architecture remains the same, and one FC layer and a softmax layer are added to let the network learn domain-specific features.

Layer 6: This layer has 512 neurons, and is fully connected to both the last FC layer in the model, which has 512 dimensions, and gender embedding of that particular image, followed by a ReLU and dropout layer.

Layer 7: This layer is a softmax layer considered as our classifier output, and contains eight neurons (classes) for the age classifier connected to the last FC layer. Due to technical details of softmax function, we can consider the output vector of this layer as the probability of each class being true.

There are a few variants in regard to choosing the proper gender embedding for each image (each of last two FC layers in gender classification model). In addition, we proposed two architectures for age classification to incorporate the gender embedding in our cross-learning. In our experiments, we tried both architectures, to evaluate the prediction accuracy.

A schematic of this architecture can be seen in Figure 3 and Figure 4.

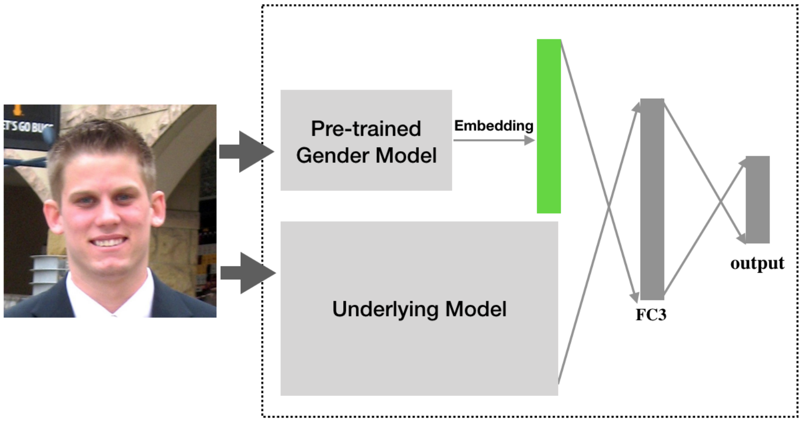

Fig. 3. Illustration of our first architecture (Manual 1). We use the FC layer of our pre-trained gender classifier as the embedding of the input, and fed it into FC3 of our prediction head along with FC2 of our underlying model

End-to-End (E2E) Architecture. Following the work of [69] by Google, we also followed the inception model, in order to give the model flexibility to combine all variants, and choose the embedding and concatenation layer which works best in the age classifier (out of four possibilities here). To this end, we design an architecture in which the network considers all variations, and the last FC layer sets proper weights for each variation, and outputs the predicted label. From the performance perspective, most computations remain shared, and all convolutional layers remain unchanged. Therefore, the number of parameters does not grow significantly. The detailed structure could be observed in Figure 5. The rest is the same as the architecture explained formerly.

4.3 Training

For re-training the gender classifier, weights are initialized randomly from a Gaussian distribution with zero mean and standard variation 0.01 at all layers. Then, for the age classification model, weights are initialized from the pre-trained gender classifier, since the purpose of prediction head is to make the

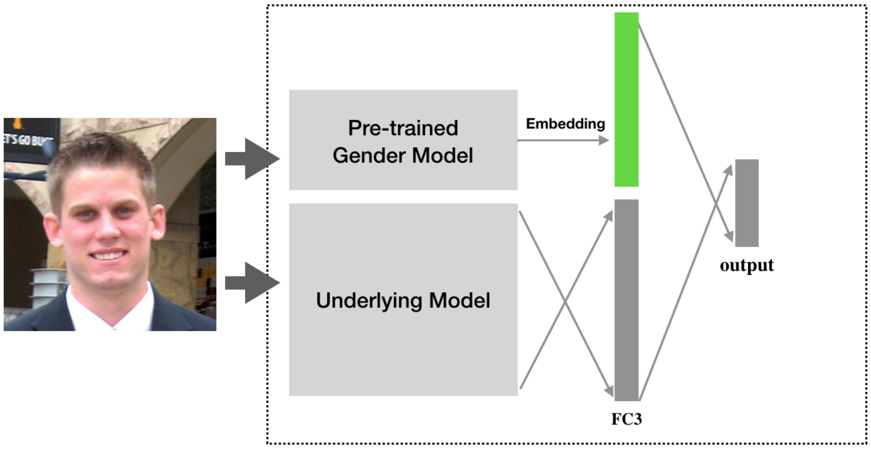

Fig. 4. Illustration of our second architecture (Manual 2). We use the FC layer of our pre-trained gender classifier as the embedding of the input, and fed it into softmax layer of our prediction head along with FC3 of our underlying model. This design gives less room to the model to fine-tune

model domain-specific. Also, we use cross-entropy as our loss function. We train the model without using any data outside of the dataset. Target values are represented as sparse binary vectors corresponding to the ground truth classes, in which the correct class is represented as one, and elsewhere is zero. For each task, the target vector is of size eight (for age classification) or two (for gender classification).

We deploy two methods to limit the risk of over-fitting. First, we use dropout. Based on the architecture, we have two dropout layers, each with a dropout ratio of 0.5 (50% chance of setting a neuron to zero). Second, we use data augmentation, by taking a random crop of 227 × 227 from the image, and randomly mirroring it in each forward-backward training pass.

Training is performed using stochastic gradient descent with batch size of 50 images, and learning rate of e^{-3}, reduced to e^{-4} after $10,000$ iterations.

4.4 Testing

We experiment with two methods for cropping the input image of size 256×256 as below. Figure 6, shows the cropping approaches.

- Center-Crop: Cropping 227×227 pixels from the center around the face, and then feeding the network.
- Over-Sampling: Input image is cropped five times. Four 227×227 pixels from each corner, and one 227×227 from the center is cropped, and all five 227×227 cropped images along with their horizontal reflections are fed to

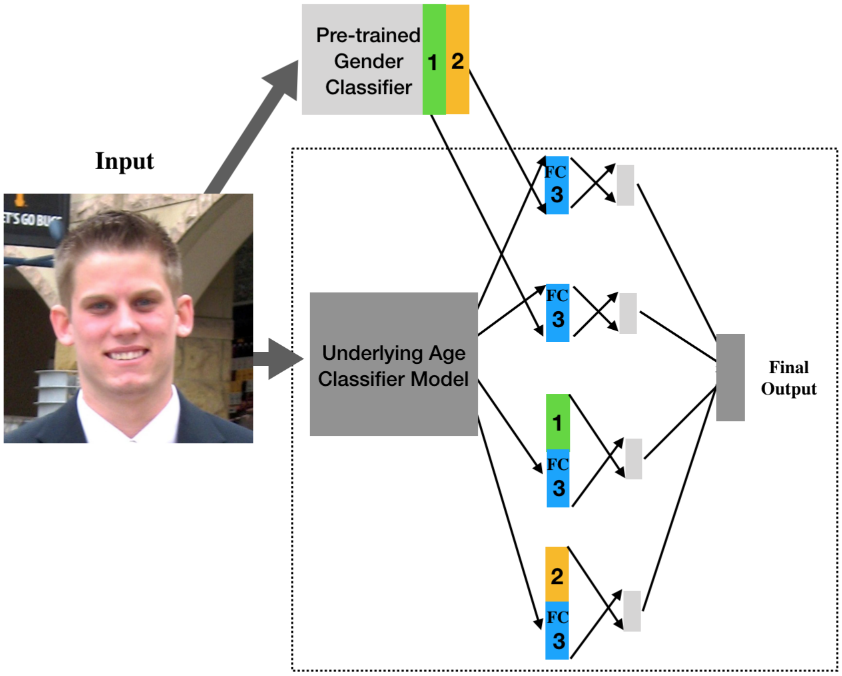
Pre-trained Gender Classifier
1
2
Input
FC 3
FC 3
Underlying Age Classifier Model
1
FC 3
2
FC 3
Final Output

Fig. 5. Illustration of the E2E system. This diagram represents how automation works in our design. We are assuming the gender classifier has two FC layers followed by each other. We call them 1 and 2 and use the colors green and yellow, respectively. In the age classifier, after having output of the convolutional layers, we split it into four copies with regard to different variations

the network, and the final prediction would be the average prediction for all the variations.

In fact, small misalignments due to challenging nature of Adience dataset can negatively impact the final results. However, the second over-sampling method is designed to compensate for these small misalignments by feeding the network with multiple translated versions of the same face.

4.5 Iterative Model Refinement

So far, we have kept the gender classifier unchanged as it was proposed by [1]. To this end, we have achieved an improvement in the age classifier, which outperforms the current state-of-the-art methods.

Similarly, we can build the gender classifier with the same approach upon the newly trained age classifier. Therefore, we achieve a better gender classification compared to what we have currently had. In our age classification approach, the accuracy of the system highly depends on the embedding of the gender,

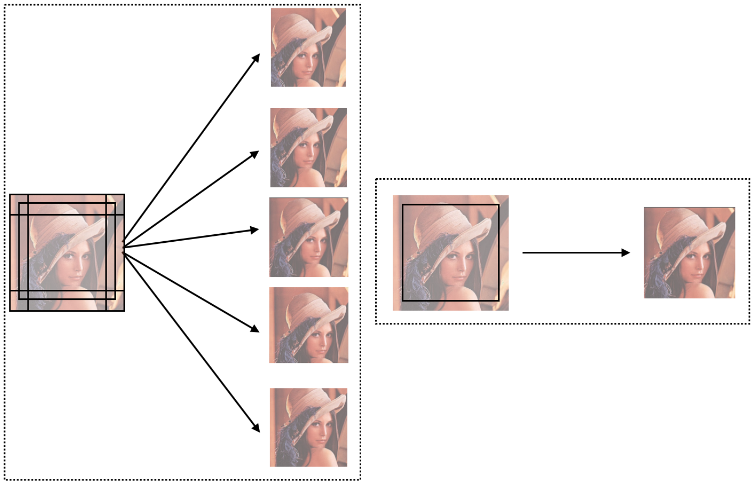

Fig. 6. Two cropping methods. Left design, shows over-sampling method. For a random image (256 × 256), we crop five squares of size 227 × 227. This method improves alignment qualities. Right design, shows center-crop method, which for any random image (256 × 256), it crops the center of size 227 × 227 around the face

and consequently the accuracy of the gender classifier. Therefore, having a more accurate gender classifier helps to re-train the age model more reliably.

As a result, we can keep repeating the last two steps, which basically provides better classifiers. Essentially, each time we re-train age (gender) from scratch, since we have more accurate gender (age) classifier by now, we are expecting to train a better age (gender) classifier. Ultimately, at some point, the system improvement stops. Intuitively, this indicates the system cannot infer more useful information from the embeddings. However, due to the limited size of dataset, and limited variations among the pictures, we have not been able to achieve promising results with this technique, but we strongly believe that having a richer dataset would provide better results for both age and gender classifiers. This approach can be seen in Figure 7.

5 System and Evaluation

Our system is implemented in Tensorflow [70], an open-source framework supported by Google, due to its advantages in parallelism, simplicity, and rich resources. We trained the network on a GeForce GTX 1060 GPU with 6 GBs of memory.

Our gender classifier takes roughly five hours to be trained. Also, our age classification model takes about 10 hours to be trained. There are two reasons why it takes longer than the gender classification architecture. First, we have one more layer, which leads to more parameters to be trained at each step.

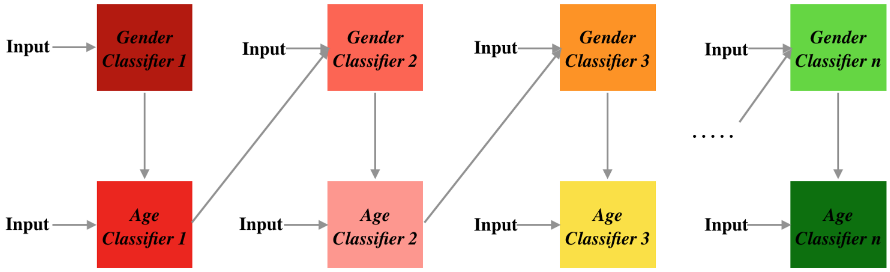

Fig. 7. Iterative Model Refinement Approach. This diagram represents the idea of iterative model refinement, when we have two tasks, which improvement in one can also improve the other task's accuracy. We have initially the gender classifier one pre-trained, hence we can build age classifier one upon that, and have improvement over first age classifier accuracy, and then we can repeat this as long as we have improvement over one or both tasks

Table 1. Adience dataset breakdown by gender category

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-	Total
Male	745	928	934	734	2308	1294	392	442	8192
Female	682	1234	1360	919	2589	1056	433	427	9411
Both	1427	2162	2294	1653	4897	2350	825	869	19487

In addition, for each image batch, we need to first pass it through the gender classifier in order to get the embedding of each image first, and then use the embedding to feed to the FC layer during each step of training. At test time, each image takes $300ms$ on average to be processed and the system outputs predicted label.

5.1 Adience Benchmark

We train and test the accuracy of our deep model on the newly released Adience benchmark [71]. The Adience dataset mostly contains images which automatically were uploaded to Flickr. Since these images were uploaded by mobile devices with no filters, viewing images are highly unconstrained. It is worth mentioning that as opposed to other datasets like LFW collection [17], which images are taken with filters on media web pages or social websites, Adience dataset images are challenging in their nature, due to their extreme variations in pose, occlusion, lighting condition, and other factors.

Adience dataset contains roughly $26K$ images of $2,284$ subjects. The breakdown of each category can be seen in table 1. Testing for both age and gender classification is performed using a standard five-fold, subject-exclusive, cross validation protocol defined in [71]. In the next section, we are comparing the previously reported results with the new results our design has produced.

5.2 Results

In table 2, we compare the results of our gender classifier which is the same as the results by [1] since we followed their architecture. Moreover, in table 3, our results for different designs can be observed. In addition, in table 4, our iterative model refinement approach results can be seen, which is not as expected due to limited variation in dataset for genders. For our age classifier, we measure the accuracy of the system both when the design predicts the exact age group, and when the system prediction is different from the correct category by at most one (correct category, or two adjacent categories) which we call it "1-off". This follows others who have done this in the past.

Table 2. Gender classification methods and accuracies

Method	Accuracy
Best from [71]	77.8 ± 1.3
Best from [72]	79.3 ± 0.0
Best from [1] single-crop	85.9 ± 1.4
Best from [1] over-sample	86.8 ± 1.4

Both proposed age and gender classifiers have been compared with methods described in [71], and [1], since Adience dataset is the benchmark which we tested our systems. Also, the proposed gender classifier has been compared to results by [72], which uses the same gender classification pipeline of [71] applied to more effective alignment of the faces.

Table 3. Age classification methods and accuracies (over-sampling)

Method	Exact	1-off
Best from [71]	45.1 ± 2.6	79.5 ± 1.4
single-crop[1]	49.5 ± 4.4	84.6 ± 1.7
over-sample[1]	50.7 ± 5.1	84.7 ± 2.2
Manual 1 Architecture	53.3 ± 3.2	85.9 ± 2.5
Manual 2 Architecture	55.1 ± 2.6	88.5 ± 2.9
E2E System	60.2 ± 1.8	92.5 ± 1.8

Based on our results which can be seen in table 3, we outperform the most accurate age classifiers with considerable gaps on Adience benchmark. We point out that we use over-sampling method in all of our experiments, since it performs better compared to the center crop method.

Table 4. Model refinement iterations and accuracies

Classifier	Step	Accuracy
Gender	1	86.8 ± 1.4
Age	1	55.1 ± 2.6
Gender	2	84.8 ± 1.2

6 Conclusions

Age and gender classifiers still lack accuracy. In this work, we developed a set of techniques that leverage a gender classifier, to improve the accuracy of age classification. We evaluated several variants on the Adience dataset, achieving an improvement of over 9% compared to the current techniques. In future work, we plan to investigate further optimization of the gender, age, and other cross-model learning, for instance, iteratively refining the models from one embedding to another.

References

1. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops. (June 2015)
2. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (June 2014) 1701–1708
3. Howard, D.: Is a man's skin really different? The International Dermal Institute
4. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. Trans. Info. For. Sec. **9**(12) (December 2014) 2170–2179
5. Mäkinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3) (2008) 541–547
6. Reid, D., Samangooei, S., Chen, C., Nixon, M., Ross, A.: Soft biometrics for surveillance: An overview (01 2013)
7. Golomb, B.A., Lawrence, D.T., Sejnowski, T.J.: SEXNET: A neural network identifies sex from human faces. In: Advances in Neural Information Processing Systems 3, [NIPS Conference, Denver, Colorado, USA, November 26-29, 1990]. (1990) 572–579
8. O'Toole, A.J., Vetter, T., Troje, N.F., Bülthoff, H.H.: Sex classification is better with three-dimensional head structure than with image intensity information. Perception **26**(1) (1997) 75–84 PMID: 9196691.
9. Moghaddam, B., Yang, M.: Learning gender with support faces. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5) (2002) 707–711
10. Baluja, S., Rowley, H.A.: Boosting sex identification performance. International Journal of Computer Vision **71**(1) (2007) 111–119
11. Toews, M., Arbel, T.: Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. IEEE Trans. Pattern Anal. Mach. Intell. **31**(9) (2009) 1567–1581
12. Chen, J., Shan, S., He, C., Zhao, G., Pietikäinen, M., Chen, X., Gao, W.: WLD: A robust local image descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9) (2010) 1705–1720
13. Ullah, I., Aboalsamh, H., Hussain, M., Muhammad, G., Mirza, A., Bebis, G.: Gender recognition from face images with local lbp descriptor. **65** (09 2012) 353–360
14. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image Vision Comput. **16**(5) (1998) 295–306
15. Perez, C., Tapia, J., Estevez, P., Held, C.: Gender classification from face images using mutual information and feature fusion. International Journal of Optomechatronics **6**(1) (1 2012) 92–119
16. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters **33** (2012) 431–437
17. B. Huang, G., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. (10 2008)
18. Akbulut, Y., Şengür, A., Ekici, S.: Gender recognition from face images with deep learning. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). (Sept 2017) 1–4
19. Mansanet, J., Albiol, A., Paredes, R.: Local deep neural networks for gender recognition. Pattern Recognition Letters **70** (2016) 80–86

20. Antipov, G., Berrani, S., Dugelay, J.: Minimalistic cnn-based ensemble model for gender prediction from face images. Pattern Recognition Letters **70** (2016) 59–65
21. Zhang, K., Tan, L., Li, Z., Qiao, Y.: Gender and smile classification using deep convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016. (2016) 739–743
22. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11) (2010) 1955–1976
23. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: Human vs. machine performance. In: International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain. (2013) 1–8
24. Salvador, A., Hynes, N., Aytar, Y., Marín, J., Ofli, F., Weber, I., Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. (2017) 3068–3076
25. Kwon, Y.H., da Vitoria Lobo, N.: Age classification from facial images. In: Conference on Computer Vision and Pattern Recognition, CVPR 1994, 21-23 June, 1994, Seattle, WA, USA. (1994) 762–767
26. Ramanathan, N., Chellappa, R.: Modeling age progression in young faces. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA. (2006) 387–394
27. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12) (Dec 2007) 2234–2240
28. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Processing **17**(7) (2008) 1178–1188
29. Fu, Y., Huang, T.S.: Human age estimation with regression on discriminative aging manifold. IEEE Trans. Multimedia **10**(4) (2008) 578–584
30. INRIA: The fg-net aging database. available: www-prima.inrialpes.fr/fgnet/html/benchmarks.html (2002)
31. Jr., K.R., Tesafaye, T.: MORPH: A longitudinal image database of normal adult age-progression. In: Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG 2006), 10-12 April 2006, Southampton, UK. (2006) 341–345
32. Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., Huang, T.S.: Regression from patch-kernel. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. (2008)
33. Fukunaga, K.: Introduction to statistical pattern recognition. (1991) 1–592
34. Yan, S., Liu, M., Huang, T.S.: Extracting age information from local spatially flexible patches. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA. (2008) 737–740
35. Ghahramani, Z.: An introduction to hidden markov models and bayesian networks. IJPRAI **15**(1) (2001) 9–42
36. Zhuang, X., Zhou, X., Hasegawa-Johnson, M., Huang, T.: Face age estimation using patch-based hidden markov model supervectors. In: 2008 19th International Conference on Pattern Recognition. (Dec 2008) 1–4

37. Gao, F., Ai, H.: Face age classification on consumer images with gabor feature and fuzzy LDA method. In: Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings. (2009) 132–141
38. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Processing **11**(4) (2002) 467–476
39. Guo, G., Mu, G., Fu, Y., Dyer, C.R., Huang, T.S.: A study on automatic age estimation using a large database. In: IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009. (2009) 1986–1991
40. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2** (1999) 1019–1025
41. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12) (2006) 2037–2041
42. Choi, S.E., Lee, Y.J., Lee, S.J., Park, K.R., Kim, J.: Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognition **44**(6) (2011) 1262–1281
43. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3) (1995) 273–297
44. Chao, W., Liu, J., Ding, J.: Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recognition **46**(3) (2013) 628–641
45. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA. (2003) 11–18
46. He, X., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]. (2003) 153–160
47. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Computer Vision - ECCV'98, 5th European Conference on Computer Vision, Freiburg, Germany, June 2-6, 1998, Proceedings, Volume II. (1998) 484–498
48. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. (2009) 256–263
49. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: The first manually collected, in-the-wild age database. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017. (2017) 1997–2005
50. Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision **126**(2-4) (2018) 144–157
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. CoRR **abs/1409.0575** (2014)
52. Pei, W., Dibeklioglu, H., Baltrusaitis, T., Tax, D.M.J.: Attended end-to-end architecture for age estimation from facial expression videos. CoRR **abs/1711.08690** (2017)

53. Chen, J., Kumar, A., Ranjan, R., Patel, V.M., Alavi, A., Chellappa, R.: A cascaded convolutional neural network for age estimation of unconstrained faces. In: 8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016. (2016) 1–8
54. Xing, J., Li, K., Hu, W., Yuan, C., Ling, H.: Diagnosing deep learning models for high accuracy age estimation from a single image. Pattern Recognition **66** (2017) 106–116
55. Liu, H., Lu, J., Feng, J., Zhou, J.: Group-aware deep feature learning for facial age estimation. Pattern Recognition **66** (2017) 82–94
56. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G.: Recent advances in convolutional neural networks. CoRR **abs/1512.07108** (2015)
57. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1**(4) (1989) 541–551
58. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. (2012) 1106–1114
59. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. (2014) 1653–1660
60. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. (2012) 2480–2487
61. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013. (2013) 3476–3483
62. Wu, Y., Hassner, T.: Facial landmark detection with tweaked convolutional neural networks. CoRR **abs/1511.04031** (2015)
63. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. (2017) 3691–3700
64. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. CoRR **abs/1611.00851** (2016)
65. Dehghan, A., Ortiz, E.G., Shu, G., Masood, S.Z.: DAGER: deep age, gender and emotion recognition using convolutional neural network. CoRR **abs/1702.04280** (2017)
66. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. (2013) 6645–6649
67. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. (2014) 1725–1732
68. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on

Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. (2017) 4236–4244

69. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR **abs/1409.4842** (2014)
70. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D.G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P.A., Vanhoucke, V., Vasudevan, V., Viégas, F.B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. CoRR **abs/1603.04467** (2016)
71. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Trans. Information Forensics and Security **9**(12) (2014) 2170–2179
72. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. (2015) 4295–4304