Image and Vision Computing 77 (2018) 10–20

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Learning CNNs from weakly annotated facial images☆

Vojtěch Franc*, Jan Čech

Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

ARTICLE INFO

Article history:
Received 16 September 2017
Received in revised form 6 April 2018
Accepted 26 June 2018
Available online 2 July 2018

Keywords:
Convolution neural networks
EM algorithm
Face recognition
Age and gender prediction
Weak annotations

ABSTRACT

Learning of convolutional neural networks (CNNs) to perform a face recognition task requires a large set of facial images each annotated with a label to be predicted. In this paper we propose a method for learning CNNs from weakly annotated images. The weak annotation in our setting means that a pair of an attribute label and a person identity label is assigned to a set of faces automatically detected in the image. The challenge is to link the annotation with the correct face. The weakly annotated images of this type can be collected by an automated process not requiring a human labor. We formulate learning from weakly annotated images as a maximum likelihood (ML) estimation of a parametric distribution describing the weakly annotated images. The ML problem is solved by an instance of the EM algorithm which in its inner loop learns a CNN to predict attribute label from facial images. Experiments on age and gender estimation problem show that the proposed algorithm significantly outperforms the existing heuristic approach for dealing with this type of data. A practical outcome of our paper is a new annotation of the IMDB database [26] containing 300 k faces each one annotated by biological age, gender and identity labels.

1. Introduction

Convolutional neural networks (CNNs) learned from examples achieve the state-of-the-art performance in many face recognition problems. Achieving good performance however requires a large set of facial images annotated by an attribute label to be predicted. Annotation of large image databases is laborious. A prototypical application addressed in this paper is the age and gender prediction. While the facial images are abundant on the Internet the biological age of captured subjects is not easily accessible and a possible manual annotation is costly and imprecise. The publicly available databases are of limited size and very often contain specific distribution of faces. For example, the two most frequently used databases, the FG-NET [23] and the MORPH [25], contain 1002 and 55,000 faces, respectively. Moreover, the MORPH database is composed of images of criminal suspects with significantly biased apparent age if compared to a normal population.

A possible solution is to create the annotation by an automated process. For example, Rothe et al. [26] created a database with 524,230 images of celebrities downloaded from imdb.com and Wikipedia. The crawler also downloaded a profile information like the person's name, the gender and the year of birth. The age was subsequently calculated as a difference between the photo taken date available in EXIF and the year of birth that is known for the celebrities. This process annotates each database image by person's name, biological age and gender. Faces in the images are found automatically by a face detector which often returns multiple detections in a single image. An example of a weakly annotated image is shown in Fig. 1. The authors of Ref. [26] use a simple heuristic to associate the annotations with the detected faces. The images with a single or a dominant face detection are assumed to contain the target person. This process creates a database of 260,282 facial images labeled with age and gender which is far more than has any other existing public dataset. The IMDB + WIKI database annotated by this heuristic is the core component of the current state-of-the-art method, e.g. it was used by winners of ChaLearn Looking At People 2015 competition [10] as well as by winners [3] of the follow up competition ChaLearn LAP 2016 and the recently published works, e.g. Refs.[1,21]. Since a significant portion of images is mislabeled, all mentioned works use the IMDB + WIKI dataset only for pre-training weights of a CNN which is followed by a fine-tuning on a smaller dataset with precise annotation.

In this paper, we propose a principle method for learning CNNs from weakly annotated images. We assume that each database image is assigned a pair of an attribute label and an identity label

☆ This paper has been recommended for acceptance by Vitomir Štruc.

* Corresponding author.
E-mail addresses: xfrancv@cmp.felk.cvut.cz (V. Franc), cechj@cmp.felk.cvut.cz (J. Čech).

https://doi.org/10.1016/j.imavis.2018.06.011
0262-8856/

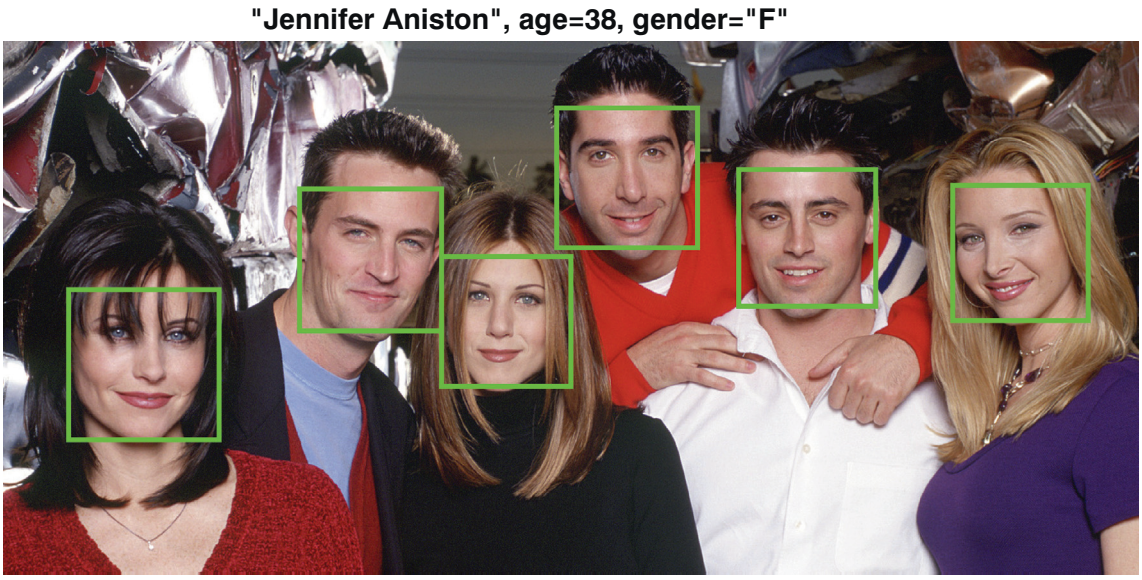


Fig. 1. An example of a weakly annotated image from the IMDB database. Each database image is assigned the identity, the biological age and the gender of a person which should appear among the faces detected in the image. The challenge is to link the annotation with the correct face.

corresponding to a single out of possibly many persons detected in the image. We further assume that each identity appears in multiple images from the database. The IMDB database of Ref. [26] is a special instance of the weakly annotated database in which the attribute label encodes gender and biological age. Our method is however generic and it can deal with other attributes as well.

This paper presents the following contributions:

1. We define a statistical model describing the distribution of the weak annotations. The integral part of our model are two CNNs describing relation between face images, attribute labels and identity labels.
2. We show how to learn parameters of the model from weakly annotated images by regularized Maximum-Likelihood method. We use the posterior regularization method of Ref. [12] to enforce the constraint on the number of failures of the used face detector. The learning problem is solved by an instance of the EM algorithm [9,27] which has two main outputs. The first output is a CNN for prediction of the attribute label from an arbitrary face image. The second output is a database of fully annotated face images. In particular, each face image is annotated by unique attribute label and identity label.
3. We applied the proposed method to learn a CNN for age and gender prediction from the weakly annotated IMDB database. The achieved prediction accuracy significantly outperforms the CNN trained from the same images annotated by the existing heuristic method of Ref. [26].
4. Unlike annotation heuristic of Ref. [26], the proposed method does not rely on images with a single face detection. We experimentally verified that removing the single detections from the IMDB database has a negligible impact on the prediction accuracy when the proposed method is used while the heuristic method becomes inapplicable.
5. The proposed method annotates 300 k faces from the IMDB database by age, and gender and identity. We used a subset of manually annotated images to verify that the automatically generated annotation is correct in more than 92% of cases. In contrast, the heuristic of Ref. [26] which selects less then 200 k faces out of which only 80% are correctly annotated.

This paper extends our previous work published in Ref. [11] by the following five improvements. First, our previous method required a small database of fully annotated faces for EM initialization while the new method works solely with weakly annotated images. Second, the new method uses more complex identity model which significantly improves the accuracy of the assignment of faces to the identities. Third, we endow the likelihood method by a novel regularization term which allows to exploit a prior knowledge on the number of failures of the used face detector. Fourth, a balance of the attribute (age and gender) and identity label cues for detection-annotation assignment was achieved by adjusting the softmax distribution of the attribute label. Fifth, testing is performed by using challenging cross-database experiments on major publicly available datasets. The proposed method outperforms the existing heuristic approach in both age and gender prediction and detection-annotation assignment, while the old version was superior for age and gender prediction only.

Most existing works related to automatic age estimation (and estimation of soft-biometrics in general) use supervised learning methods, for example, Refs.[5,14,15,20] etc. The supervised methods require fully annotated examples, that is, pairs of facial image and a single attribute label. Learning age estimation from weakly annotated faces has been addressed scarcely. Most existing works in this category assume that the training set contains pairs of facial image and a weak attribute label. For example, instead of an exact age, like in supervised methods, a weak label can be an interval of admissible ages [4,30] or age distribution [13]. In general, learning classifiers from ambiguously labeled examples (also known as learning from partially annotated examples) has been attacked by various approaches including e.g. risk minimization methods [8,18], Expectation Maximization methods [19], dictionary methods [7,32] or matrix completion [6]. These methods consider a scenario when each input instance is annotated by a set of candidate labels only one of which is known to be correct. The setting addressed in our paper is different and it can be seen as a generalization of the multi-class multi-instance learning (MIL) [2,28,31]. The multi-class MIL assumes that the training instances are grouped into bags and the labels are assigned to the bags rather than to individual instances. The main assumption is that the bag label is correct at least for

one instance in the bag. Zeng et al. [31] proposes a variant of MIL which allows the bags to be annotated by a subset of identities that appear in the image. Our work extends the multi-class MIL setting to the case when each bag is assigned a pair of labels each of different nature, namely, we consider the attribute label and the identity label. The existing MIL methods are not directly applicable because a naive merging of the two labels into a single one would lead to an intractably large number of classes. Another major difference is that our method allows to use learnable CNNs for image representation while the existing works like Refs. [28,31] rely on a prescribed set of features.

2. Statistical model of weakly annotated images

In this paper, we address the problem of learning from weakly annotated images. A weakly annotated image depicts a scene with possibly multiple human faces when at most one of them belongs to the person for which we have an annotation. The annotation describes the person's identity, his/her age and gender. The faces in the image are located automatically by a face detector. It is unknown which of the detected faces belong to the annotation. Moreover, it is possible that the annotated person is not among the detected faces which happens if the person is not present in the image or if the face detector fails. An additional important property of the data we consider is that each identity must appear in multiple images. An example of weakly annotated image from the IMDB database is shown in Fig. 1.

In order to exploit the data for learning of face recognition systems we have to establish the missing link between the annotation (age, gender, identity) and faces detected in the images. The proposed method attempts the problem by simultaneously learning models of identity, age and gender which are then used to annotate the images. Learning of the models and the annotation of images is done in alternating fashion. The result of the proposed method is an annotated database, identity models of the annotated persons and prediction models that allow to estimate age and gender from an arbitrary facial image.

In the reminder of this section, we first define exactly the concept of weakly annotated images and then we describe a statistical model governing their distribution. The statistical model depends on multiple parameters. Learning of the model parameters from data is a subject of Section 3.

2.1. The weakly annotated training set

The weakly annotated training set $\mathcal{T} = \left\{\left(\mathbf{X}^j, y^j, c^j\right) \in \mathcal{X}^* \times \mathcal{Y} \times \mathcal{C} \mid j = 1, \ldots, m\right\}$ contains m triplets, each describing a single weakly annotated image. The tensor $\mathbf{X}^j = \left(\mathbf{x}^j_1, \ldots, \mathbf{x}^j_{n^j}\right) \in \Re^{100\times100\times n_j}$ represents a bag of n^j facial images extracted from the j-th image. The facial images are cropped around boxes found by a face detector. The pre-processing of the facial images involves i) re-sizing them to 100×100 pixels, ii) re-scaling the pixel intensities to the range 0–255 and iii) subtracting the mean image computed from all facial images detected in the database. The symbol $y^j \in \mathcal{Y}$ denotes the attributed label of the annotated person that should be captured in the j-th image. In our case, the label encodes the person's gender and age restricted to the range 16–75 years, that is,

$$\mathcal{Y} = \{\text{male\_16}, \text{male\_17}, \ldots, \text{male\_75}, \text{female\_16}, \text{female\_17}, \ldots, \text{female\_75}\}.$$

The symbol $c^j \in \mathcal{C} = \{1, \ldots, C\}$ denotes the identity of the annotated person where C is the total number of identities in the database. The identity labels are obtained by mapping each name string to a unique integer from 1 to C.

We assume that the training set $\mathcal{T}$ contains samples drawn from random variables which are independently and identically distributed according to

$$p_{\boldsymbol{\theta}}(\mathbf{X}, y, c) = p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, c)\, p(\mathbf{X}, c).$$

The symbol $\boldsymbol{\theta}$ denotes a vector composed of all model parameters. In the sequel, we decompose the distribution $p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, c)$ into several components and describe them separately. On the other hand, the distribution $p(X, c)$ is not modeled and estimated in the course of the learning algorithm.

2.2. Latent variables

For each weakly annotated image $(\mathbf{X}, y, c)$, we introduce an auxiliary vector of latent binary variables $\mathbf{h} = (h_1, \ldots, h_n) \in \mathcal{H} = \left\{\mathbf{h}' \in \{0,1\} \mid\mid \mathbf{h}'\|_1 \leq 1\right\}$[1]. The latent variables determine which of the faces $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ corresponds to the annotation (y, c). The value $\mathbf{h} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the vector with 1 at the i-th position and zeros everywhere else, means that the annotation (y, x) belongs to the image $\mathbf{x}_i$. The value $\mathbf{h} = \mathbf{0}$ means that the annotated identity is not among the detected faces $\mathbf{X}$. We assume that the annotated identity appears in the image at most once and hence the admissible latent vectors must satisfy the condition $\|\mathbf{h}\|_1 \leq 1$. Analogously to $(\mathbf{X}, y, c)$, the vector $\mathbf{h}$ is also assumed to be a realization of a random variable.

We further assume that the attribute label y and the identity label c are conditionally independent given $\mathbf{X}$ and $\mathbf{h}$, that is,

$$p(y, c \mid \mathbf{X}, \mathbf{h}) = p(y \mid \mathbf{X}, \mathbf{h})\, p(c \mid \mathbf{X}, \mathbf{h}). \tag{1}$$

The assumption Eq. (1) means that all information about the attribute label y is contained in solely the facial image determined by $(\mathbf{X}, \mathbf{h})$. The same holds for the identity label c. Note that the assumption would be violated if we had additional knowledge constraining the relation between y and c, for example, the death age of the identity c. Under the assumption Eq. (1), we can decompose $p_{\boldsymbol{\theta}}(y, h \mid \mathbf{X}, c)$ as a product

$$p_{\boldsymbol{\theta}}(y, h \mid \mathbf{X}, c) = p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, \mathbf{h})\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, c). \tag{2}$$

Consequently, we can obtain $p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, c)$, governing the distribution of data in the training set $\mathcal{T}$, by marginalizing out the latent variable $\mathbf{h}$, that is,

$$p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, c) = \sum_{\mathbf{h}\in\mathcal{H}} p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, \mathbf{h})\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, c). \tag{3}$$

The Eq. (3) decomposes the distribution of weakly annotated images into two parts: the model of the attribute label $p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, \mathbf{h})$ and the identity model $p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, c)$. The two parametric distributions are described in the following subsections.

2.3. The model of attribute labels

Given faces $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ detected in a single image and the corresponding latent variables $\mathbf{h} = (h_1, \ldots, h_n)$, the attribute label y is distributed according to

$$p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, \mathbf{h}) = \begin{cases} p(y) & \text{if } \mathbf{h} = \mathbf{0}, \\ p_{\boldsymbol{\theta}}(y \mid \mathbf{x}_i) & \text{if } \mathbf{h} = \mathbf{e}_i, \end{cases}$$

[1] Later, we use for each example $(\mathbf{X}^j, y^j, c^j)$ a different set of latent variables, denoted as $\mathcal{H}^j = \{1, \ldots, n^j\}$, because the number of detected faces n^j is not the same for all images.

meaning that if the annotated person is not among the faces, $\mathbf{h} = \mathbf{0}$, then the attribute label is distributed according to a prior distribution $p(y)$. Otherwise, when the i-th face corresponds to the annotated person, $\mathbf{h} = \mathbf{e}_i$, then the attribute label distribution reads

$$p_{\boldsymbol{\theta}}(y \mid \mathbf{x}_i) = \alpha + (1 - \alpha|\mathcal{Y}|)\frac{\exp(\langle \mathbf{v}_y, \boldsymbol{\psi}(\mathbf{x}_i)\rangle)}{\sum_{y' \in \mathcal{Y}} \exp(\langle \mathbf{v}_{y'}, \boldsymbol{\psi}(\mathbf{x}_i)\rangle)},$$

where $\boldsymbol{\psi}(\mathbf{x}) \in \mathbb{R}^{2048}$ are features extracted from $\mathbf{x}$ by a CNN and $\mathbf{v}_y \in \mathbb{R}^{2048}$, $y \in \mathcal{Y}$, are parameters of its last layer. The hyper-parameter $\alpha \in \left[0, \frac{1}{|\mathcal{Y}|}\right]$ controls the minimal probability assigned to any label from $\mathcal{Y}$. For $\alpha = 0$, the probability $p_{\boldsymbol{\theta}}(y \mid \mathbf{x}_i)$ is computed by passing the image $\mathbf{x}_i$ through the CNN with the soft-max distribution as the last layer. The standard soft-max distribution allows the probability mass to be concentrated around a single label while the remaining labels have probability close to zero. This case often happens in the early iterations of EM algorithm that is used to learn the parameters. As a result, the faces with attribute label probability close to zero are effectively excluded from consideration. This problem is healed by setting $\alpha > 0$ which prevents the EM algorithm to trust the current estimate of the attribute label probability with too high confidence. The value of α effectively controls the influence of the attribute (age and gender) model when assigning the annotation to the detected faces in the E-step of the EM algorithm. In our experiments, we use the value $\alpha = 0.005$ which was found by tuning the model on a small subset of manually annotated training images.

The convolution filters of the CNN, defining the feature extractor $\boldsymbol{\psi}$, as well as the weights $(\mathbf{v}_y, y \in \mathcal{Y})$ of the penultimate layer are encapsulated in the parameter vector $\boldsymbol{\theta}$ and they are learned in the course of the EM algorithm. A detailed configuration of the CNN is described in Table 1.

2.4. The identity model

Given faces $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ detected in a single image and the corresponding label c determining identity of the annotated person, the latent variables $\mathbf{h} = (h_1, \ldots, h_n)$ are distributed according to

$$p_\theta(\mathbf{h} \mid \mathbf{X}, c) = \begin{cases} \omega/Z & \text{if } \mathbf{h} = \mathbf{0}, \\ s(\| \boldsymbol{\phi}(\mathbf{x}_i) - \mathbf{w}_c \|)/Z & \text{if } \mathbf{h} = \mathbf{e}_i, \end{cases}$$

Table 1
Configuration of the CNN used to predict label from a facial image. The second column describes the number of filters 'filt', the filter size 'k', stride 's' and padding 'p'.

Layer type	Configuration
Output	Distribution over Y outputs
Soft-Max	
Convolution	filt: Y, k: 1 × 1, s: 1, p: 0
ReLU	
Convolution	filt: 2048, k: 1 × 1, s: 1, p: 0
ReLU	
Convolution	filt: 2048, k: 5 × 5, s: 1, p: 0
ReLU	
Convolution	filt: 128, k: 4 × 4, s: 1, p: 0
ReLU	
Convolution	filt: 128, k: 3 × 3, s: 1, p: 0
MaxPool	2 × 2, s: 2, p: 0
ReLU	
Convolution	filt: 64, k: 3 × 3, s: 1, p: 0
MaxPool	2 × 2, s: 2, p: 0
ReLU	
Convolution	filt: 64, k: 3 × 3, s: 1, p: 0
MaxPool	2 × 2, s: 2, p: 0
ReLU	
Convolution	filt: 32, k: 3 × 3, s: 1, p: 0
ReLU	
Convolution	filt: 32, k: 3 × 3, s: 1, p: 0
Input	100 × 100 gray-scale image

where $Z = \omega + \sum_{i=1}^{n} s(\| \boldsymbol{\phi}(\mathbf{x}_i) - \mathbf{w}_c \|)$ is the normalization constant, ω is a parameter to be learned and

$$s(d) = \frac{1}{1 + \exp(\gamma(d^2 - \beta^2))}$$

is a smooth approximation of the step-function with the hyper-parameters $\gamma > 0$ and $\beta > 0$. The symbol $\boldsymbol{\phi}(\mathbf{x}_i) \in \mathbb{R}^{4096}$ denotes the identity descriptor extracted from the image $\mathbf{x}_i$. In our experiments, the descriptor is L_2-normalized output of the penultimate layer of the VGG-Face CNN [24]. For our setting of the hyper-parameters, the value of $s(d)$ is approximately 1 for d in the interval $[0, \beta)$, it is equal to 0.5 for $d = \beta$, and it is close to 0 for $d > \beta$. Therefore, the value of $p_\theta(\mathbf{h} = \mathbf{e}_i \mid \mathbf{X}, c)$, expressing the probability the face $\mathbf{x}_i$ belongs to the identity c, is large if the descriptor $\boldsymbol{\phi}(\mathbf{x}_i)$ lies inside the ball with radius β and centered in the template $\mathbf{w}_c$. If the descriptor $\boldsymbol{\phi}(\mathbf{x}_i)$ is outside that ball the probability quickly drops to zero, and how quickly it drops it is controlled by the hyper-parameter γ. Finally, the parameter ω corresponds the probability of the event that none of the faces belongs to the identity. Fig. 2 visualizes the function $s(d)$ and meaning of the hyper-parameters.

In our experiments, we used $\gamma = 20$ and $\beta = 0.5$ which were obtained by tuning the model on a small subset of manually labeled training images. We also used the posterior regularization method to enforce the value of ω/Z to be 0.15 which corresponds to the portion of images where the detector failed to find the celebrity. The identity templates $(\mathbf{w}_c, c \in \mathcal{C})$ are contained in the parameter vector $\boldsymbol{\theta}$ and they are learned in the course of the EM algorithm together with other parameters. Note that unlike the CNN for extraction of the attribute label, the parameters of the VGG-Face descriptor are not learned although it would be possible in principle. However, we found the pre-trained VGG-Face CNN sufficiently good.

3. Learning the model parameters

In this section, we describe a method for learning parameters $\boldsymbol{\theta}$ of the distribution Eq. (3) from the weakly annotated training set $\mathcal{T}$. To measure the match between the model parameters $\boldsymbol{\theta}$ and the data $\mathcal{T}$, we define the conditional log-likelihood

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{m} \log p_\theta(y^j \mid \mathbf{X}^j, \mathbf{c}^j) = \sum_{j=1}^{m} \log \sum_{\mathbf{h} \in \mathcal{H}^j} p_\theta(y^j \mid \mathbf{X}^j, \mathbf{h}) p_\theta(\mathbf{h} \mid \mathbf{X}^j, c^j).$$

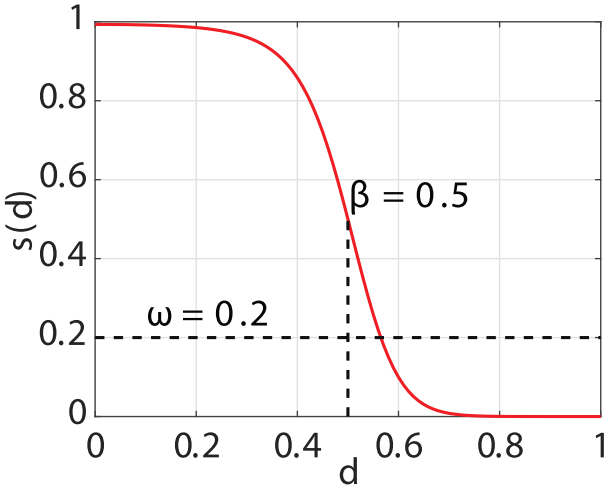


Fig. 2. Visualization of $s(d) = 1/(1 + exp(\gamma(d^2 - \beta^2)))$ which we use as a smooth approximation of the step-like function in the definition of the identity model. The "steepness" parameter γ is set to 20.

Besides the data $\mathcal{T}$, the learning algorithm is desired to exploit a strong prior knowledge about the posterior distribution of latent variables $p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)$. In particular, we know that the detector fails to find the annotated person in a small portion of images. The event of the detector's failure in the j-th image has the probability $p_{\boldsymbol{\theta}}(\mathbf{h} = \mathbf{0} \mid \mathbf{X}^j, y^j, c^j)$. Therefore, we want the posterior probability to satisfy

$$\frac{1}{|\mathcal{J}_c|} \sum_{j \in \mathcal{J}_c} p_{\boldsymbol{\theta}}(\mathbf{h} = \mathbf{0} \mid \mathbf{X}^j, y^j, c^j) \leq \tau, \qquad \forall\, c \in \mathcal{C}, \tag{4}$$

where $\mathcal{J}_c = \{j \in \{1, \ldots, m\} \mid y^j = c\}$ are the indices of images associated with identity c, τ is the maximal portion of images where the identity is not among the detected faces. Using a small portion of manually inspected images we found that the probability of the detector failure is around 15%, that is, $\tau = 0.15$ (more details how this value is estimated can be found in Section 6.1).

We are not aware of a tractable algorithm that would be able to maximize the likelihood $L(\boldsymbol{\theta})$ subject to the constraint Eq. (4). Instead, we use a tractable method proposed in Ref. [12] which enforces Eq. (4) via a regularization term included in the objective function. Let $\mathcal{Q}$ denote a set of all posterior distributions satisfying the constrain Eq. (4). We measure the distance between the set $\mathcal{Q}$ and the posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)$ by

$$D(\boldsymbol{\theta}) = \min_{q \in \mathcal{Q}} \mathrm{KL}(q(\mathbf{h}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)),$$

where $\mathrm{KL}(\cdot \| \cdot)$ denotes the Kullback-Leibler divergence. Our learning algorithm then finds the parameters $\boldsymbol{\theta}$ by maximizing a regularized log-likelihood, that is,

$$\hat{\boldsymbol{\theta}} \in \operatorname*{Argmax}_{\boldsymbol{\theta}} \quad [L(\boldsymbol{\theta}) - D(\boldsymbol{\theta})]. \tag{5}$$

The learning algorithm based on Eq. (5) pursues two goals simultaneously: i) fitting the model to data by maximizing $L(\boldsymbol{\theta})$ and ii) considering only those models whose posterior is close to the set $\mathcal{Q}$ by minimizing the distance $D(\boldsymbol{\theta})$. In the next section, we describe an efficient algorithm finding a local maximum of Eq. (5). We found experimentally that learning without the posterior regularization leads to significantly worse results. Moreover, implementing the regularization requires only a slight modification of the algorithm as will be seen bellow.

3.1. The algorithm

The problem Eq. (5) can be reformulated as

$$\hat{\boldsymbol{\theta}} \in \operatorname*{Argmax}_{\boldsymbol{\theta}} \quad \max_{q \in \mathcal{Q}} F(\boldsymbol{\theta}, q) \tag{6}$$

where

$$F(\boldsymbol{\theta}, q) = L(\boldsymbol{\theta}) - \mathrm{KL}(q(\mathbf{h}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)).$$

The equivalence between problems (5) and (6) follows from the identity $L(\boldsymbol{\theta}) - D(\boldsymbol{\theta}) = \max_{q \in \mathcal{Q}} F(\boldsymbol{\theta}, q)$. A natural algorithm to solve the problem (6) is then the block-coordinate ascent alternating maximization w.r.t. variables $\boldsymbol{\theta}$ and q.

It is interesting to consider a special case when the set $\mathcal{Q}$ does not restrict the posterior distribution, that is when τ in Eq. (4) equals to 1. In this case, $\max_q F(\boldsymbol{\theta}, q) = L(\boldsymbol{\theta})$ because $\min_q \mathrm{KL}(q(\mathbf{h}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)) = 0$. As a result optimization problem (6) is equivalent to maximization of the log-likelihood $L(\boldsymbol{\theta})$ without any constraint on the posterior distribution. The block-coordinate ascent algorithm maximizing $F(\boldsymbol{\theta}, q)$ is then known as the EM algorithm [9,27]. Hence we use the same name for the algorithm solving Eq. (6) in the constrained case as well. The method is outlined in Algorithm 1.

Algorithm 1. EM-algorithm solving Eqs. (5) and (6)

1: $q_0 \in \mathcal{Q}$.
2: **repeat**
3: M-step: $\theta_t = \operatorname{argmax}_\theta F(\theta, q_{t-1})$
4: E-step: $q_t = \operatorname{argmax}_{q \in \mathcal{Q}} F(\theta_t, q)$
5: **until** convergence.

The algorithm starts from an initial distribution $q_0 \in \mathcal{Q}$. In our experiments, we set $q_0^j(\mathbf{h} = \mathbf{0}) = \tau$ and $q_0^j(\mathbf{h} = \mathbf{e}_i) = \frac{1-\tau}{n^j}$, $i \in \{1, \ldots, n^j\}$ for all $j \in \{1, \ldots, m\}$. This means that at the beginning all detected faces has an equal chance to be the annotated identity and the probability of detector's failure is in each image equal to τ. The convergence of the algorithm can be assessed by monitoring the objective function $F(\boldsymbol{\theta}_t, q_t)$.

The maximization of $F(\boldsymbol{\theta}, q)$ w.r.t. $\boldsymbol{\theta}$ and q is in the literature denoted as the M-step and the E-step, respectively. The M-step and the E-step can be decomposed into a set of simpler problems after re-writing the objective $F(\boldsymbol{\theta}, q)$ as follows:

$$\begin{aligned}
F(\boldsymbol{\theta}, q) &= L(\boldsymbol{\theta}) - \mathrm{KL}\left(q(\mathbf{h}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c)\right) \\
&= \sum_{j=1}^{m} \log p_{\boldsymbol{\theta}}(y^j \mid \mathbf{X}^j, c^j) + \sum_{j=1}^{m} \sum_{h \in \mathcal{H}^j} q^j(\mathbf{h}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}^j, y^j, c^j)}{q^j(\mathbf{h})} \\
&= \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log p_{\boldsymbol{\theta}}(y^j \mid \mathbf{X}^j, c^j) p(\mathbf{h} \mid \mathbf{X}^j, y^j, c^j) - \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log q^j(\mathbf{h}) \\
&= \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log p_{\boldsymbol{\theta}}(y^j \mid \mathbf{X}^j, \mathbf{h}) p(\mathbf{h} \mid \mathbf{X}^j, c^j) - \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log q^j(\mathbf{h}).
\end{aligned}$$

All identities are obtained just by rearrangement of appropriate terms. The last equation uses the identity $p_{\boldsymbol{\theta}}(y \mid \mathbf{X}, c) p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}, y, c) = p_{\boldsymbol{\theta}}(y, h \mid \mathbf{X}, c)$ and the formula (2). In the next sections, we describe how to solve the E-step and M-step efficiently.

3.2. Solving the M-step

The M-step involves maximization of $F(\boldsymbol{\theta}, q_{t-1})$ w.r.t. the model parameters $\boldsymbol{\theta}$. We split the model parameters into two parts $\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_b)$. The part $\boldsymbol{\theta}_A$ encapsulates the prior probability $p(y)$, the convolution filters $\boldsymbol{\psi}$ of the CNN modeling the attribute labels and attribute parameters $(\mathbf{v}_y, y \in \mathcal{Y})$ of the CNN's last layer. The part $\boldsymbol{\theta}_B$ encapsulates the templates $(\mathbf{w}_c, c \in \mathcal{C})$ of the identity model over the VGG-Face descriptor. We then write the function $F(\boldsymbol{\theta}, q_{t-1})$ as a sum

$$F(\boldsymbol{\theta}, q_{t-1}) = F_A(\boldsymbol{\theta}_A, q_{t-1}) + F_B(\boldsymbol{\theta}_B, q_{t-1}) + F_C(q_{t-1}),$$

where

$$F_A(\boldsymbol{\theta}_A, q_{t-1}) = \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log p_{\boldsymbol{\theta}_A}(y^j \mid \mathbf{X}^j, \mathbf{h}),$$

$$F_B(\boldsymbol{\theta}_B, q_{t-1}) = \sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log p_{\boldsymbol{\theta}_B}(\mathbf{h} \mid \mathbf{X}^j, c^j),$$

$$F_C(q_{t-1}) = -\sum_{j=1}^{m} \sum_{\mathbf{h} \in \mathcal{H}^j} q^j(\mathbf{h}) \log q^j(\mathbf{h}).$$

This decomposition allows us to decompose the M-step into maximization of $F_A(\boldsymbol{\theta}_A, q_{t-1})$ w.r.t. $\boldsymbol{\theta}_A$ and $F_B(\boldsymbol{\theta}_B, q_{t-1})$ w.r.t. $\boldsymbol{\theta}_B$ independently. The last term, $F_C(q_{t-1})$, does not depend on the parameters and hence it can be ignored. Solution of the two sub-problems is discussed below.

3.2.1. Update of the attribute label model

In this section, we discuss maximization of $F_A(\boldsymbol{\theta}_A, q_{t-1})$ w.r.t. $\boldsymbol{\theta}_A$ encapsulating parameters of the attribute label model. By exploiting the particular form of $p_{\boldsymbol{\theta}_A}(y^j \mid \mathbf{X}^j, \mathbf{h})$ we can rewrite the objective function as

$$F_A(\boldsymbol{\theta}, q_{t-1}) = \sum_{j=1}^{m} q_{t-1}^j(\mathbf{h} = \mathbf{0}) \log p(y^j) + \sum_{j=1}^{m} \sum_{i=1}^{n^j} q_{t-1}^j(\mathbf{h} = \mathbf{e}_i) z\left(\frac{\exp(\langle \mathbf{v}_{y^j}, \boldsymbol{\psi}\left(\mathbf{x}_i^j\right)\rangle)}{\sum_{y' \in \mathcal{Y}} \exp(\langle \mathbf{v}_{y'}, \boldsymbol{\psi}\left(\mathbf{x}_i^j\right)\rangle)} \right),$$

where $z(t) = \log(\alpha + (1 - \alpha|\mathcal{Y}|)t)$. It is seen that the maximization of $F_A(\boldsymbol{\theta}_A, q_{t-1})$ can be done for the prior probability $p(y)$ and the CNN parameters $(\boldsymbol{\psi}, \mathbf{v}_y, y \in \mathcal{Y})$ independently. First, the new value of prior $p(y)$ is obtained by solving

$$\max_{p(y)} \sum_{j=1}^{m} q_{t-1}^j(\mathbf{h} = \mathbf{0}) \log p(y^j) \quad \text{s.t.} \quad \sum_{y \in \mathcal{Y}} p(y) = 1, p(y) \geq 0, y \in \mathcal{Y}. \tag{7}$$

The problem Eq. (7) is convex and its closed form solution is derived from the Karush-Kuhn-Tucker (KKT) conditions, in particular,

$$p(y) = \frac{\sum_{j=1}^{m} [\![y^j = y]\!] q_{t-1}^j(\mathbf{h} = \mathbf{0})}{\sum_{j=1}^{m} q_{t-1}^j(\mathbf{h} = \mathbf{0})}, \qquad y \in \mathcal{Y},$$

where $[\![A]\!]$ is the Iverson bracket. Second, the CNN parameters $(\boldsymbol{\psi}, \mathbf{v}_y, y \in \mathcal{Y})$ are updated by solving

$$\max_{\boldsymbol{\psi}, \mathbf{v}_y, y \in \mathcal{Y}} \sum_{i=1}^{m} \sum_{i=1}^{n_j} q_{t-1}^j(\mathbf{h} = \mathbf{e}_i) z\left(\frac{\exp\left(\left\langle \mathbf{v}_{y^j}, \boldsymbol{\psi}\left(\mathbf{x}_i^j\right)\right\rangle\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\left\langle \mathbf{v}_{y'}, \boldsymbol{\psi}\left(\mathbf{x}_i^j\right)\right\rangle\right)} \right). \tag{8}$$

Note that replacing $z(t) = \log(\alpha + (1 - \alpha|\mathcal{Y}|)t)$ by arbitrary strictly monotonically increasing function will not change the maximizers of the problem Eq. (8). For instance, if we use $z(t) = \log(t)$ the objective of Eq. (8) becomes weighted average of the soft-max log-loss. Hence, updating the parameters $(\boldsymbol{\psi}, \mathbf{v}_y, y \in \mathcal{Y})$ boils down to standard supervised learning of a classification CNN when each training example $\left(\mathbf{x}_i^j, y^j\right)$ has a weight $q_{t-1}^j(\mathbf{h} = \mathbf{e}_i)$. Recall, that $q_{t-1}^j(\mathbf{h} = \mathbf{e}_i)$ means the probability of $\mathbf{x}_i^j$ being the face of the person the attribute y^j belongs to.

3.2.2. Update of the identity model

In this section, we discuss maximization of $F_B(\boldsymbol{\theta}_B, q_{t-1})$ w.r.t. $\boldsymbol{\theta}_B$ encapsulating the templates $(\mathbf{w}_c, c \in \mathcal{C})$ of the identity model. Using the particular form of $p_{\boldsymbol{\theta}_B}(\mathbf{h}|\mathbf{X}, c)$, we can write the objective function as $F_B(\boldsymbol{\theta}_B, q_{t-1}) = \sum_{c \in \mathcal{C}} \sum_{j \in \mathcal{J}_c} F_B^c(\mathbf{w}_c, q_{t-1}) + K$ where

$$F_B^c(\mathbf{w}_c, q_{t-1}) = \sum_{j \in \mathcal{J}_c} \left[\sum_{i=1}^{n_j} q_{t-1}^j(\mathbf{h} = \mathbf{e}_i) \log s\left(\left\| \mathbf{x}_i^j - \mathbf{w}_c \right\| \right) - \log\left(\sum_{i=1}^{n_j} s\left(\left\| \mathbf{x}_i^j - \mathbf{w}_c \right\| \right) + \tau \right) \right]$$

and $K = \sum_{j=1}^{m} q_{t-1}^j(\mathbf{h} = 0) \log \tau$ is a constant independent of $\mathbf{w}_c$. Hence, each template $\mathbf{w}_c, c \in \mathcal{C}$, can be updated independently by solving an unconstrained problem $\max_{\mathbf{w}_c} F_B^c(\mathbf{w}_c, q_{t-1})$. Since the objective $F_B^c(\mathbf{w}_c, q_{t-1})$ is differentiable w.r.t $\mathbf{w}_c$ everywhere we can use arbitrary algorithm for smooth optimization.

3.3. Solving the E-step

The E-step problem involves maximization of $F(\boldsymbol{\theta}_t, q)$ w.r.t the distribution $q \in \mathcal{Q}$ which can be decomposed into C (the number of identities) independent problems of the form

$$\max_{q} \sum_{j \in \mathcal{J}_c} \sum_{\mathbf{h} \in \mathcal{H}^j} \left(q^j(\mathbf{h}) \log A^j(\mathbf{h}) - q^j(\mathbf{h}) \log q^j(\mathbf{h}) \right) \tag{9a}$$

subject to

$$\begin{aligned} \textstyle\sum_{\mathbf{h} \in \mathcal{H}} q^j(\mathbf{h}) &= 1, \qquad j \in \mathcal{J}_c, \\ q^j(\mathbf{h}) &\geq 0, \qquad j \in \mathcal{J}_c, \mathbf{h} \in \mathcal{H}^j, \\ \frac{1}{|\mathcal{J}_c|} \textstyle\sum_{j \in \mathcal{J}_c} q^j(\mathbf{h} = \mathbf{0}) &\leq \tau, \end{aligned} \tag{9b}$$

where $\mathcal{J}_c = \{j \in \{1, \ldots, m\} \mid y^j = c\}$ are indices of images associated to the identity c and $A^j(\mathbf{h}) = p_{\boldsymbol{\theta}_t}(y^j \mid \mathbf{X}^j, \mathbf{h}) p_{\boldsymbol{\theta}_t}(\mathbf{h} \mid \mathbf{X}^j, c^j)$ is a shortcut. Since the objective function is strictly concave and the constrains form a convex polyhedron, the problem Eq. (9a) is convex and it has a unique solution. Any off-the-shelf convex solver can be used to solve Eq. (9a), however, the particular form of the problem allows to use a simple algorithm which is outlined next.

From the Karush-Kuhn-Tucker optimality conditions, we learn that the optimal solution of Eq. (9a) has the form

$$q^j(\mathbf{h} = \mathbf{0}) = \frac{A^j(\mathbf{h} = \mathbf{0}) \exp(-\lambda)}{Z^j} \text{ and } q^j(\mathbf{h} = \mathbf{e}_i) = \frac{A^j(\mathbf{h} = \mathbf{e}_i)}{Z^j}, \; j \in \mathcal{J}_c, \tag{10}$$

where $Z^j = A^j(\mathbf{h} = \mathbf{0}) \exp(-\lambda) + \sum_{i=1}^{n^j} A^j(\mathbf{h} = \mathbf{e}_i)$ and $\lambda \geq 0$ is a non-negative constant. The constant λ is found by solving a univariate problem obtained after substituting Eq. (10) to Eq. (9a). The problem boils down to finding the maximal λ satisfying the constraint

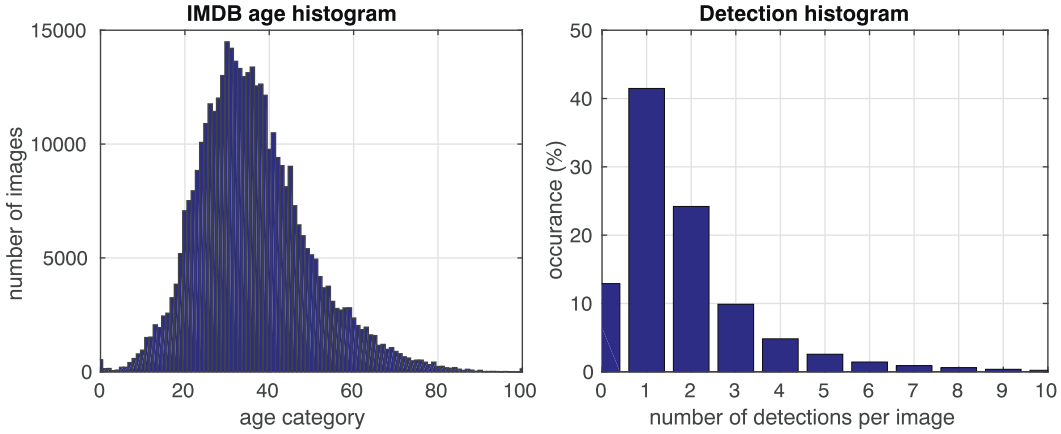
$$\frac{1}{|\mathcal{J}_c|} \sum_{j \in \mathcal{J}_c} \frac{A^j(\mathbf{h} = \mathbf{0}) \exp(-\lambda)}{A^j(\mathbf{h} = \mathbf{0}) \exp(-\lambda) + \sum_{i=1}^{n_j} A^j(\mathbf{h} = \mathbf{e}_i)} \leq \tau, \tag{11}$$

which can be solved efficiently by the binary search.

4. Datasets

4.1. IMDB dataset

The dataset collected by Rothe et al. [26] consists of 460,723 images of celebrities (mainly actors) downloaded from imdb.com. The crawler also downloaded a profile information, so beside a person's name a year of birth, and a gender was stored. The age was subsequently calculated as the difference between the photo taken date from EXIF tag and the year of birth. This process is not error free. There are minor cases of apparently incorrect age (negative age due to wrong EXIF tag, age over 100 due to scanned photos (George Washington 300 years, J.F. Kennedy 90 years, Jack London 134 years), or due to name coincidence (Kathryn Boyd 110 years). We discarded all images having age out of the range [16, 75] or invalid gender. See distribution of age categories of IMDB dataset in Fig. 3 (a).



(a) Age distribution. (b) Distribution of detected faces.

Fig. 3. Subfigure (a) shows a distribution of age categories in the IMDB dataset and Subfigure (b) a histogram of number of detected faces per image.

In many cases, multiple persons appear on the image. The dataset is not distributed with any detections or bounding boxes of the target person. We ran a commercial multi-view detector[2] on the entire dataset. This resulted in 859,198 detections. A single face is found in about 41% of images. No detection occurs in 13% of images and the remaining 46% of images contain multiple detections, see Fig. 3(b).

We created two training sets that are used in the following evaluation. The full training set is created from the IMDB database by taking all images which contain at least one face detection and have age annotation inside the range [16, 75]. The reduced training set is created from the full training set by discarding images that contain a single face detection. Table 2 shows the basic statistics of the original IMDB database and the two training sets used in the evaluation.

4.2. Test datasets

We use a very challenging cross-database protocol for testing the age-gender prediction accuracy. Note that most existing works on age-gender recognition construct the training and the test set by randomly splitting images from the same database. The latter case is significantly less difficult because the training and the test images are from the very same distribution.

Our test set includes several publicly available datasets: APPA-REAL [1], ChaLearnAge [10], FG-NET [23], and LFW [17] comprising 19,253 images in total. See Table 3 for details. The number of test images results from restricting the original examples to those with age in [16–75] range. Note that more than half of our test images captures ordinary people and not the celebrities who may appear younger due to a professional make up or even aesthetic surgery. Needless to say that the imaging quality differs among the datasets. We believe, mostly these factors make the cross-dataset experiment particularly challenging.

5. Baselines

In this section, we describe two engineering solutions to the problem of selecting the detections from the IMDB dataset so that they represent the target persons in as many cases as possible.

[2] Courtesy of Eyedea Recognition, Ltd. www.eyedea.cz

5.1. Baseline 1: dominant face detection

In the first approach to the selection problem, we follow Ref. [26]. All single detections are taken. Additionally, for images where the second strongest detection is below a threshold ($\tau_{2nd} = 70$), the strongest detection is also taken if it is above a threshold ($\tau_{1st} = 130$). The thresholds are set empirically based on tuning on a small set of manually annotated images. The minimum detection score to output a bounding box is 30. The second condition adds about 7 k images. The *baseline-1* creates a dataset containing 187,211 annotated faces.

The selection heuristic is based on the assumption that the target person is present in all the single detections and that a dominant detection belongs to the target person. None of the assumption always holds, since a target person may be missed by the detector while the other person on the same image is detected or the target person may have lower detection score than other person due to expression, pose, occlusion or lighting condition.

5.2. Baseline 2: median identity from single detections

We propose a novel heuristic selection strategy which exploits the annotation of the person identity. For all detections, an identity descriptor $\boldsymbol{\phi}(\mathbf{x})$ is computed. The descriptor is 4096-dimensional

Table 2
The number of images the number of detected faces shown for the original IMDB database and the two training set used in evaluation. The full training set contains all images with age in the range [16, 75] which have at leas one detection. The reduced set is created from the full set by discarding images with one detection.

	IMDB database	Full training set	Reduced training set
Number of images	460,723	377,198	196,447
Number of detected faces	859,198	801,594	620,843

Table 3
Test datasets. The datasets used in our cross-dataset experiments for testing. The number of images results from restricting the original examples to those with age in the range [16, 75].

Dataset	Num. of images	Attributes	Content	Annotation
APPA-REAL [1]	6077	Age	Ordinary people	Public
ChaLearnAge [10]	3490	Age + gender	Ordinary people	Public
FG-NET [23]	421	Age + gender	Ordinary people	Public
LFW [17]	9265	Age + gender	Celebrities	Ours
All	19,253			

L_2-normalized output of the penultimate layer of the VGG-Face CNN [24]. A single etalon is calculated for each celebrity by computing a component-wise median over all single detections. It is assumed that the majority of single detections is correct. Then, for every image in the IMDB dataset, a detection that is the closest to the etalon of the target celebrity is selected, if L_2 distance between the detection and the median is below empirically set threshold $\tau_{id} = 0.9$. Note that images of celebrities without any single detections are not considered and some of the single detections may be rejected. The *baseline-2* creates a dataset containing 307,153 annotated faces.

6. Experiments

6.1. Implementation details

We implemented the proposed method, i.e. the EM-CNN Algorithm 1, in Matlab. The EM-CNN decomposes the learning problem (6) into several independent optimization sub-tasks. As it is shown in Sections 3.2 and 3.3, some of the sub-tasks have closed form solution while others have to be solved numerically. To this end, we used the following optimization methods:

- The update of the CNN for the attribute prediction leads to solving a minimization task Eq. (8). In our experiments, we solve the problem approximately by applying 20 epochs of the SGD algorithm with momentum implemented in MatConvNet toolbox [29]. Since solving the problem Eq. (8) constitutes the most computationally demanding part of the M-step, we apply this parameter update only every 5th iteration of the EM algorithm. Note, that skipping updates in some iterations does not affect convergence of the EM to a local optimum.
- The optimization task connected with the update of the identity model, as described in Section 3.2.2, is solved by L-BFGS algorithm [22] implemented in C. The L-BFGS converges in a fraction of second for problems with hundreds of images per identity.
- The E-step boils down to a constrained optimization problem Eq. (9a) which can be solved efficiently in its dual form. The dual task requires to find a maximal λ satisfying the constraint (11) which can be achieved by a binary search. We found that 100 divisions of the interval $\gamma \in [0, 10]$ always led to a very accurate solution obtained in a fraction of second. Alternatively, the original problem Eq. (9a) can be attacked by any off-the-shelf convex solver.

The proposed method has four hyper-parameters that must be selected manually prior to running the EM algorithm:

- The identity model $p_\theta(\mathbf{h} \mid \mathbf{X}, c)$ has two hyper-parameters $\gamma > 0$ and $\beta > 0$. We run the EM algorithm on a small subset of IMDB faces (extracted from 2000 images associated with 200 identities) for different values of γ and β. The learned model was then used to assign the annotation to detected faces. The prediction error was computed based on 3352 manually annotated images. Based on the prediction error we found the best setting of the hyper-parameters to be $\gamma = 20$ and $\beta = 0.5$.
- The appearance model $p_\theta(y \mid \mathbf{x})$ has a single hyper-parameter $\alpha \in \left[0, \frac{1}{|\mathcal{Y}|}\right]$ which controls the influence of the attribute (age and gender) when assigning the annotation to the detected faces in the image. Using the same approach as for tuning γ and β, except that we minimized the age-gender prediction error in this case, we found the best value to be $\alpha = 0.005$.
- The hyper-parameter $\tau \in [0, 1]$ is an upper bound on the probability that the detector does not find the annotated identity in a randomly selected image. The detector failure is of two kinds. First, the annotated identity appears in the image but the face detector misses the face. Second, the face of the identity is not visible in the image (e.g., the person is turned back or not captured in the image at all). We used 1440 manually annotated images to estimate that in $199/1440 \approx 14\%$ of cases the identity is not among the faces detected in the image and hence we set the upper bound to $\tau = 0.15$. We also found that in $143/199 \approx 72\%$ of cases, corresponding to 10% of all images, the identity face is visible in the image but it is missed by the face detector. In case of using a different face detector or a different image database, the value of τ needs to re-estimated.

The experiments run on a Linux machine with 128 GB RAM and Tesla K40C GPU 12 GB/3004 MHz. In all experiments, we run 40 iterations of the EM-CNN algorithm after which the likelihood function stopped improving. Learning from the full IMDB database containing ≈800 K faces required around 14 days.

6.2. Accuracy of age and gender estimation

The accuracy of the trained network was measured by three statistics: Mean Absolute Error (MAE), which is the average absolute deviation between the predicted age and the ground-truth age computed over the test set; Cumulative Score at 5 (CS5), which is a percentage of test images having the prediction error less or equal to five years; and Gender Error (gerr), which is a male-female misclassification rate. We also compute the confidence intervals of the statistics estimated on the test sample using the Hoeffding's inequality [16][3]. In particular, at the probability 95%, the confidence interval for MAE is ±0.58 years and for CS5 and gerr it is ±0.98%.

We used two baseline CNNs learned from fully annotated faces which were selected by the heuristics *baseline-1* (187,211 faces) and *baseline-2* (307,153 faces) on the IMDB dataset. We used the proposed algorithm, denoted as EM-CNN, to learn a CNN from the weakly annotated images. In order to make learning of the identity model stable, we removed images corresponding to identities with less than 5 example images. This reduction left us with 764,625 faces (originally 801,594) in 360,977 images (originally 377,198).

The *baseline-1*, *baseline-2* and the EM-CNN learn networks with the same configuration of the layers (c.f. Table 1). The distribution $p_{\boldsymbol{\theta}}(y \mid \mathbf{x})$, with $\boldsymbol{\theta}$ learned by one of the three competing methods, is used to construct plug-in Bayes predictors of the age and gender using the MAE and the 0/1-loss, respectively. The prediction accuracy estimated on images from the test dataset is summarized in Table 4.

It is seen that training from the smallest dataset selected by *baseline-1* has the worst error statistics. Better results are achieved by training from data collected by the proposed *baseline-2*. The proposed EM-CNN training outperforms all baselines in MAE and CS5, and is very similar in gerr to *baseline-2*. However, the difference between *baseline-2* and EM-CNN is not very significant. The results indicate that the proposed EM-CNN can handle the problem well, but the heuristic selection strategy *baseline-2*, based on a representation from single detections, turned out to be particularly efficient for IMDB dataset.

Much more challenging is a situation where single detections are not present, i.e. there are always at least two faces detected on every image in the dataset. To evaluate this case, we use the reduced training set created by discarding all images with a single face detection. We use the same *baseline-2* heuristic on the dataset without single

[3] The true value of the expected risk R is in the interval $(\hat{R}-\varepsilon, \hat{R}+\varepsilon)$ with probability $\delta = 0.95$, where $\varepsilon = \sqrt{\frac{K^2}{2l}(\log(2) - \log(1-\delta))}$, $l = 19{,}253$ is the number of test examples, $\hat{R}$ is the sample mean of the loss function computed on test examples and K is the maximal value of the loss. $K = 75 - 16 = 59$ for MAE and $K = 1$ for CS5 and gerr.

Table 4
The results. Besides the error statistics of the methods (MAE, CS5, gerr), the table shows the number of training samples of face images used by the corresponding learning algorithm. The statistics were estimated on an independent test dataset with 19,253 samples. The values of MAE are in the interval ±0.58 years and ±0.98% for CS5 and gerr with probability 95%.

Method	Number of trn. faces	MAE	CS5 [%]	Gerr [%]
Baseline-1	187,211	6.6	50.6	5.6
baseline-2	307,153	6.0	54.5	**4.4**
EM-CNN	764,625	**5.9**	**55.3**	4.5
Baseline-2 (woSingle)	96,481	7.5	44.9	6.8
EM-CNN (woSingle)	591,654	**6.2**	**53.6**	**4.8**

detections (woSingle) except for the fact that a celebrity etalon is computed as a median over all images where the celebrity is supposed to be present. In Table 4, it is seen that proposed EM-CNN (woSingle) outperforms the *baseline-2* (woSingle) by a significant margin. The results of EM-CNN (woSingle) are only slightly worse than *baseline-2* selecting examples from the full training set.

Fig. 4(a) shows a distribution of MAE per age category for methods trained on the full training set. It is seen that all methods perform similarly in the interval 16–40 years, while for ages above 40 years EM-CNN performs significantly better than the two baselines. In addition, MAE of EM-CNN is similar in a large range of ages, while MAE of the baselines change significantly with the age. A distribution of MAE for the experiment on the training set without single detections is shown in Fig. 4(b). The performance of EM-CNN is similar to the experiment on full training set, while the performance of *baseline-2* is noticeably worse.

6.3. Accuracy of identity recognition

Besides training the network from weak annotations, the proposed EM-CNN can be used to create a dataset of annotated face images. The annotation is assigned to the detected faces based on the posterior probability estimated by the algorithm. The most likely configuration of the latent variables is computed by $\mathbf{h}^j \in \text{Argmax}_{\mathbf{h}\in\mathcal{H}^j} p_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{X}^j, y^j, c^j)$. The j-th image is marked as not containing the target person if $\mathbf{h}^j = \mathbf{0}$. Otherwise, when $\mathbf{h}^j = \mathbf{e}_i$, the i-th detected faces $\mathbf{x}_i^j$ is annotated by the attribute label y^j and the identity label c^j. Using the model parameters learned by the EM-CNN, this strategy produces 311,085 annotated face while the remaining 49,892 images are marked as not containing the target identity. Note that the portion of excluded faces is around 16.04% showing that the bound 15% (hyper-parameter $\tau = 0.15$) enforced by the posterior regularization was effective. The number of annotated faces produced by *baseline-1* and *baseline-2* is 187,211 and 307,153, respectively.

In order to measure correctness of the assignment of the annotation to the detections, we have manually annotated a set of 3352 images from the IMDB dataset. Two subsets were annotated: the images with single detections only (1565 images), and the images with multiple detections (1787 images). Each of the subsets was randomly sampled such that all age and gender categories $\{F, M\} \times \{16, \ldots, 75\}$ are approximately uniformly present. A human annotator selects either one of the detected bounding box as the target person or none if the target person was not detected by the face detector. As an aid for the annotator, all detections were displayed together with tens of images found by Google querying the target person name. This task is sometimes not so easy for a human. The assignment is not always unambiguous, especially for low resolution and non-frontal views.

Having a set of N images with the ground-truth $\mathbf{h}_*^j$ and predicted $\mathbf{h}^j$ assignment, three errors were measured: the overall assignment error *err*, the precision *prec* defined as the portion of correctly annotated faces among those which were selected by the algorithm, and *recall* defined as the portion of correctly annotated faces among all faces that could be correctly annotated if one new the ground truth[4].

We present results separately for images with single detections only and with multiple detections only, see Table 5 (a) and (b) respectively. It is seen that the selection strategy *baseline-1* works satisfactorily on images with single detections where it has the assignment error 18.8% and precision 81.2%. However, *baseline-1* fails on the images with multiple detections where 88.1% of annotations are incorrectly linked with the detected faces. The proposed *baseline-2* works drastically better than *baseline-1* on both types of images, achieving the assignment error 9.3% and 16.4% on images with single and multiple detections, respectively. The annotation strategy learned by EM-CNN algorithm has overall the lowest assignment errors; 8.0% (single detections) and 10.8% (multiple detections). To make the results of *baseline-2* and EM-CNN comparable, we also added a constant to the value of $p_{\boldsymbol{\theta}}(\mathbf{h} = 0 \mid \mathbf{X}^j, y^j, c^j)$ so that the precision of both methods is the same (the results after adjustment are denoted by "EM-CNN (adj)"). It is seen that the adjusted EM-CNN has significantly higher recall if compared to *baseline-2*.

We repeated the experiments on the reduced training set which contains a subset of images with multiple face detections. The heuristic *baseline-2* (woSingle) is not working in this case; 53.9% of annotations are incorrectly linked to the faces. On the other hand, the performance of EM-CNN (woSingle) is only marginally below the results obtained on the full dataset; the assignment error on multiple detection images is 12.0% when learned from images each containing several persons.

Table 5 also shows the results of the previous version of the EM-CNN published in Ref. [11] (denoted as "EM-CNN (old)"). The previous version uses much simpler identity model which was sufficient to learn precise age/gender predictor but it was not so good in recognizing the identities. It is seen that the old version is slightly worse than the *baseline-2* on the full database but it is much better on the reduced training set. We note that similar comparison of the age/gender prediction accuracy is impossible due to a different test protocol used in the previous paper.

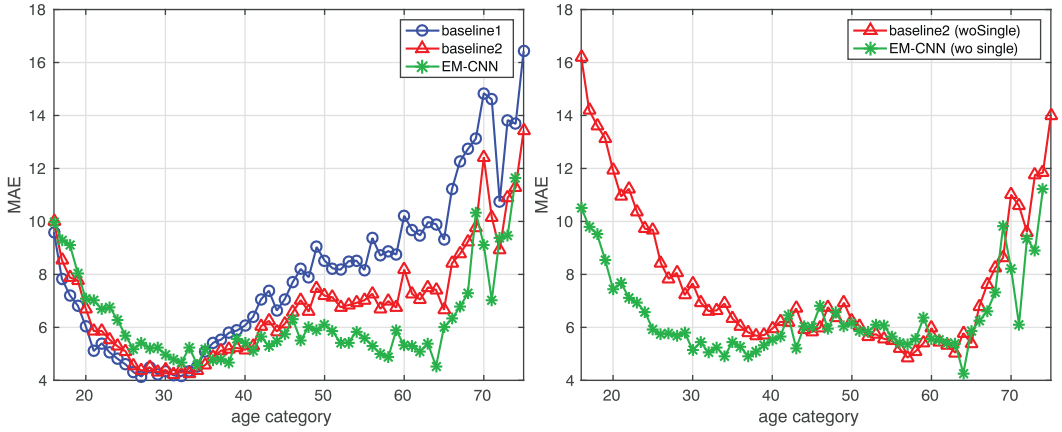
Finally, Table 6 summarizes the number of faces annotated and the precision of the annotation obtained by *baseline-1*, *baseline-2* and EM-CNN. Note that the precision on the whole dataset is computed from the precision estimated on the subsets of manually annotated images.

7. Conclusions and future work

In this paper, we have addressed a problem of learning CNNs to perform face recognition tasks from weakly annotated images. A weakly annotated image in our setting is assigned a pair of an attribute label and an identity label corresponding to a single person that should be captured in the image. It is unknown which face out of many automatically extracted faces from the image corresponds to the annotation. It is further assumed that each identity is associated with multiple images in the database.

Our main contribution is a principled approach which formulates learning from weakly annotated images as a regularized maximum likelihood estimation of a parametric distribution describing the data. The ML problem is solved by an instance of the EM algorithm which in its M-step learns a CNN for prediction of the attribute label and, in its E-step links the annotations with the faces. Experiments on IMDB database show that the proposed EM-CNN algorithm outperforms the so far used heuristic method of Rothe et al. [26] by a large margin. In addition, the EM-CNN algorithm does not require images

[4] The errors are defined as follow: $err = \sum_j [\mathbf{h}_*^j \neq \mathbf{h}^j]/N$, $prec = \sum_j [\mathbf{h}_*^j = \mathbf{h}^j \wedge \mathbf{h}^j \neq \mathbf{0}] / \sum_j [\mathbf{h}^j \neq \mathbf{0}]$ and $recall = \sum_j [\mathbf{h}_*^j = \mathbf{h}^j \wedge \mathbf{h}^j \neq \mathbf{0}] / \sum_j [\mathbf{h}_*^j \neq \mathbf{0}]$.



(a) All training images. (b) Images without single detections.

Fig. 4. Mean absolute error of the age estimate per age categories. Results for datasets derived from: the full training set (a), and the subset without images with single detections (b).

with a single detection unlike the existing heuristic which was also demonstrated experimentally.

The proposed EM-CNN algorithm links 311,085 faces from IMDB database with gender, age and identity label. This is likely to become the largest database of faces annotated by biological age, gender and the identity labels. We have quantitatively demonstrated that the label assignment accuracy is superior to previously used heuristic approach. The re-annotated database is available for download from http://cmp.felk.cvut.cz/~xfrancv/pages/emcnn.html.

There are several ways how to extend the proposed method:

- One of the limitations of the current method is the off-line trained CNN used for identity feature extraction. The method could be improved by including the identity CNN among the parameters that are learned by the EM algorithm.
- The current identity model represents each identity by a single template in the feature space. The model could be improved by using multiple templates that would allow to capture larger variation of the identity appearance.
- The main advantage of the IMDB database is that the biological age of the celebrities can be easily computed. In particular, the age is computed by subtracting the date of birth from the image capture time stored in EXIF. In case the method is applied to database of non-celebrities, it is likely that the date of birth would be hard to obtain. In this case, it would be possible to extend the method by including the unknown birth date among the latent variables of the model.

Table 5
Accuracy of label assignment to detected faces measured on a manually labeled subset of the IMDB images. The comparison is carried out separately on images with a single face detection (a) and images with multiple detections (b). The *baseline-1* proposed in Ref. [26] is compared against the proposed *baseline-2* and the EM-CNN algorithm. The results of the previous version of the EM-CNN published in Ref. [11] , which uses much simpler identity model, are denoted by EM-CNN (old). The EM-CNN (adj) corresponds to the label assignment of the EM-CNN after it was adjusted to match the precision of *baseline-2* in order to make the results of the two methods comparable. The best results for each evaluation metric are boldfaced.

Method	*Prec* [%]	*Recall* [%]	*Err* [%]
(a) Single detections only (1565 images).			
Baseline-1	81.2	**100.0**	18.8
Baseline-2	93.5	95.2	9.3
EM-CNN (old)	87.0	93.6	16.0
EM-CNN	**93.8**	96.5	**8.0**
EM-CNN (adj)	93.5	97.0	8.1
(b) Multiple detections only (1787 images).			
Baseline-1	51.7	1.8	88.1
Baseline-2	**92.0**	83.9	16.4
EM-CNN (old)	90.6	80.8	24.6
EM-CNN	91.1	**93.0**	**10.8**
EM-CNN (adj)	**92.0**	91.7	11.2
Baseline-2 (woSingle)	73.6	44.8	53.9
EM-CNN (woSingle, old)	88.9	88.3	20.2
EM-CNN (woSingle)	**91.5**	**90.2**	**12.0**

Table 6
The number of automatically annotated faces and the percentage of correct annotations (prec). The results are shown for the full training set and the reduced training set without the single detections (woSingle).

Method	Annotated faces	*Prec* [%]
Baseline-1	187,211	80.0
Baseline-2	307,153	92.7
EM-CNN	311,085	92.4
EM-CNN (adj)	308,619	92.7
Baseline-2 (woSingle)	104,616	73.6
EM-CNN (woSingle)	163,933	90.5

Acknowledgments

The authors were supported by Czech Science Foundation grants 16-05872S and P103/12/G084.

References

[1] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, R. Rothe, Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database, 12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017.
[2] S. Andrews, I. Tsochantaridis, T. Hofmann, Support Vector Machines for Multiple-Instance Learning, Proc. of Neural Information Processing Systems, 2002.
[3] G. Antipov, M. Baccouche, S.-A. Berrani, J.-L. Dugelay, Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models, CVPR workshop, Looking at People Challenge, 2016.
[4] K. Antoniuk, V. Franc, V. Hlavac, V-shaped interval insensitive loss for ordinal classification, Mach. Learn. 103 (2016) 261–283. .
[5] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal Hyperplane Ranker with Cost Sensitivities for Age Estimation, CVPR, 2011.
[6] C.H. Chen, V.M. Patel, R. Chellappa, Matrix Completion for Resolving Label Ambiguity, IEEE International Conference on Computer Vision (ICCV), 2015.
[7] Y.-C. Chen, V.-M. Patel, R. Chellapa, P.-J. Phillips, Ambiguously labeled learning using dictionaries, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2076–2088.
[8] T. Cour, B. Sapp, B. Taskar, Learning from partial labels, J. Mach. Learn. Res. 12 (2011) 1225–1261.

[9] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1) (1977) 1–38.
[10] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H.J. Escalante, I. Guyon, Chalearn 2015 apparent age and cultural event recognition: datasets and results, ICCV, ChaLearn Looking at People workshop, 2015.
[11] V. Franc, J. Čech, Face attribute learning from weakly annotated examples, in: B. Bir, H. Abdenour, J. Qiang, N. Mark, Š. Vitomir (Eds.), Proc. of International Conference on Automatic Face and Gesture Recognition Workshops, Biometrics in the Wild (BWILD), IEEE Computer Society, New York, US, 2017, pp. 933–940.
[12] K. Ganchev, J. Graça, J. Gillenwater, B. Taskar, Posterior regularization for structured latent variable models, J. Mach. Learn. Res. 11 (2010) 2001–2049.
[13] X. Geng, K. Smith-Miles, Z.H. Zhou, Facial Age Estimation by Learning from Label Distributions, Proc. of Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
[14] X. Geng, Z.H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Learn. 29 (12) (2007) 2234–2240.
[15] H. Han, C. Otto, A.K. Jain, Age Estimation from Face Images: Human vs. Machine Performance, International Conference on Biometrics (ICB), 2013.
[16] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (301) (1963) 13–30.
[17] 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.
[18] L. Jie, F. Orabona, Learning from Candidate Labeling Sets, Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Curran Associates Inc., USA, 2010, pp. 1504–1512.
[19] R. Jim, Z. Ghahramani, Learning with multiple labels, Proc. of NIPS, 2002.
[20] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, IEEE Trans. Pattern Anal. Mach. Learn. 24 (4) (2002) 442–455.
[21] S. Lapuschkin, A. Binder, K.-R. Muller, Understanding and Comparing Deep Neural Networks for Age and Gender Classification, Proceedings of the ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG), 2017.
[22] J. Nocedal, Updating quasi-Newton matrices with limited storage, Math. Comput. 35 (151) (1980) 773–782.
[23] G. Panis, A. Lanitis, N. Tsapatsoulis, T.F. Cootes, Overview of research on facial ageing using the FG-NET ageing database, IET Biom. 5 (2). (2016)
[24] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, British Machine Vision Conference, 2015.
[25] K.J. Ricanek, T. Tesafaye, MORPH: A Longitudinal Image Database of Normal Adult Age-Progression, IEEE 7th International Conference on Automatic Face and Gesture Recognition, IEEE, Southampton, UK, 2006. pp. 341–345. https://ieeexplore.ieee.org/document/1613043/.
[26] R. Rothe, R. Timofte, L. Van Gool, DEX: Deep EXpectation of apparent age from a single image, IEEE International Conference on Computer Vision Workshops (ICCVW), 2015.
[27] M. Schlesinger, A connection between learning and self-learning in the pattern recognition (in Russian), Kibernetika 2 (1968) 81–88.
[28] A. Shrivastava, V.-M. Patel, J.-K. Pillai, R. Chellappa, Generalized dictionaries for multiple instance learning, Int. J. Comput. Vis. 114 (2-3) (2015) 288–305.
[29] A. Vedaldi, K. Lenc, MatConvNet — Convolutional Neural Networks for MATLAB, Proceeding of the ACM Int. Conf. on Multimedia, 2015.
[30] S. Yan, H. Wang, X. Tang, J. Liu, T. Huang, Regression from uncertain labels and its applications to soft biometrics, IEEE Trans. Inf. Forensics Secur. 3 (4) (2008) 698–708.
[31] Z. Zeng, S. Xiao, T.-H. Jia, K. Chan, S. Gao, Y. Ma, Learning by Associating Ambiguously Labeled Images, Proc. of the International Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2013,
[32] M.-L. Zhang, F. Yu, Solving the Partial Label Learning Problem: An Instance-based Approach, Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, AAAI Press. 2015, pp. 4048–4054.