Learning CNNs for face recognition from weakly annotated images

Vojtech Franc and Jan Cech
Department of Cybernetics, Faculty of Electrical Engineering,
Czech Technical University in Prague, Czech Republic

***Abstract*— Supervised learning of convolutional neural networks (CNNs) for face recognition requires a large set of facial images each annotated with a single attribute 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 estimation of a parametric distribution describing the data. The ML problem is solved by an instance of EM algorithm which in its inner loop learns a CNN to perform given face recognition task. Experiments on age and gender estimation problem show that the proposed EM-CNN algorithm significantly outperforms the state-of-the-art approach for dealing with this type of data.**

I. INTRODUCTION

Convolutional neural networks (CNNs) learned from examples achieve the state-of-the-art performance in many face recognition problems. Achieving good performance requires a large set of facial images annotated by an attribute label to be predicted. Annotation of large image databases is often expensive. A prototypical application addressed in this paper is the age estimation. While the facial images are abundant on the internet the biological age of depicted persons is not easily accessible and its 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 [Panis et al., 2016] and the MORPH [Ricanek and Tesafaye, 2006], contain 1,002 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., 2015] 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. 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 typically returns multiple detections in a single image. An example of a weakly annotated image is shown in Figure 1. The authors of [Rothe et al., 2015] 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. Although significant portion of images are mislabeled, the overall quality of the annotation was enough to learn a CNN for age estimation winning the ChaLearn Looking At People 2015 competition [Escalera et al., 2015]. Also winners of the follow up competition, ChaLearn LAP 2016, use the same database and a similar annotation heuristic [Antipov et al., 2016].

The authors were supported by Czech Science Foundation grants 16-05872S and P103/12/G084, and by the internal CTU funding SGS17/185/OHK3/3T/13.

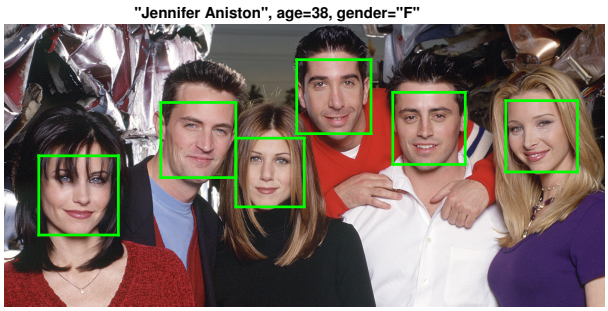


Fig. 1: An example of a weakly annotated image from the IMDB database. Each database image is assigned the identity, the biological age and the gender of a person which should appear among the faces detected in the image. The challenge is to link the annotation with the correct face.

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 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+WIKI database of [Rothe et al., 2015] is a special instance of the weakly annotated database in which the attribute label encodes a gender and a 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 parametric distribution over the attribute labels, the identity labels and the appearances of the

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.115

detected faces. An important component of the model is a CNN governing the distribution of the attribute labels conditioned on a facial image.

2) We derive an instance of the EM algorithm [Schlesinger, 1968], [Dempster et al., 1977] estimating the distribution parameters from both weakly and fully annotated images. The M-step of the EM algorithm involves training of the desired CNN which is subsequently used for prediction of attribute labels from un-annotated facial images. A byproduct are also identity models of all weakly annotated persons in the database.
3) We applied the proposed method to learn a CNN for age and gender prediction using the weakly annotated IMDB database and a medium sized fully annotated public data. The achieved prediction accuracy significantly outperforms the CNN trained from images annotated by the heuristic method of [Rothe et al., 2015].
4) Unlike the selection heuristic of [Rothe et al., 2015], the proposed method does not require 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 is not applicable.

Most existing works related to automatic age estimation (and estimation of soft-biometrics in general) use supervised learning methods, for example, [Lanitis et al., 2002], [Geng et al., 2007], [Chang et al., 2011], [Han et al., 2013] 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. The existing works assume that the training set contains pairs of facial image and weak attribute label. For example, instead of an exact age, like in supervised methods, a weak label can be an interval of admissible ages [Yan et al., 2008], [Antoniuk et al., 2016] or age distribution [Geng et al., 2010]. 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 [Cour et al., 2011], Expectation Maximization methods [Jim and Ghahramani, 2002] or matrix completion [Chen et al., 2015]. These methods consider a scenario when each input 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. It can be seen as a generalization of the multi-instance learning (MIL) [Andrews et al., 2002]. The MIL assumes that the training set is composed of bags of inputs which are collectively annotated by a single binary label. In contrast, we assume that the bag of faces is annotated by a pair of attribute label and identity label.

II. STATISTICAL MODEL OF THE DATA

Our training set is composed of a small database of fully annotated images and a large database of weakly annotated images. A fully annotated image is a facial image along with the ground-truth value of an attribute label describing the depicted person. In our case the attribute label is an integer encoding the person's biological age and gender. A weakly annotated image depicts a scene with possibly multiple faces one of them belonging to the target person. The image is annotated by the name of the target person and a label describing his/her age and gender. The database image is summarized by a set of facial images found automatically by a face detector. It is unknown which of the detected facial images contains the target person. Moreover, it is possible that the target person is not among the detected faces which happens, for example, if the detector fails or if the target persons is not visible in the scene.

In the following text we first describe a statistical model governing the distribution of the fully annotated and the weakly annotated images. Then we describe an instance of the EM algorithm for learning parameters of the statistical model using simultaneously the weakly and the fully annotated images. The statistical model provides the face descriptor of all the annotated identities in the database and a missing link between the annotation and the detected faces. The main output is a CNN learned to predict the label from an unannotated facial image.

A. FULLY ANNOTATED IMAGES

First we consider the standard supervised setting. In this case the training set $\mathscr{T}_{\mathrm{F}} = \{(\mathbf{x}^j, y^j) \in \mathscr{X} \times \mathscr{Y} \mid j = 1, \ldots, l\}$ contains l facial images $\mathbf{x}^j \in \mathscr{X}$ and their corresponding attribute labels $y^j \in \mathscr{Y}$. The symbol $\mathscr{X}$ denotes a set of all input images and $\mathscr{Y}$ is a discrete set of labels. In our experiments $\mathscr{X} = \mathbb{R}^{100\times 100}$ are normalized gray-scale images of size 100×100 pixels found by a face detector and $\mathscr{Y} = \{1, \ldots, Y\}$ encodes all combinations of a gender and a biological age of the depicted person. Hence the maximal label Y is twice the number of age categories.

We model the conditional distribution of the label y given the image $\mathbf{x}$ by

$$p_\theta(y \mid \mathbf{x}) = \frac{\exp(\langle \mathbf{v}_y, \psi(\mathbf{x})\rangle)}{\sum_{y' \in \mathscr{Y}} \exp(\langle \mathbf{v}_{y'}, \psi(\mathbf{x})\rangle)}$$

where $\psi(\mathbf{x}) \in \mathbb{R}^{2048}$ are features extracted from $\mathbf{x}$ by a CNN and $\mathbf{v}_y \in \mathbb{R}^{2048}$, $y \in \mathscr{Y}$, are parameters of its penultimate layer. The configuration of the CNN used in our experiments is detailed in Table I. Let $\theta \in \mathbb{R}^d$ be a concatenation of all convolution filters of the CNN. The parameter θ can be estimated by maximizing the conditional log-likelihood

$$L_{\mathrm{F}}(\theta; \mathscr{T}_{\mathrm{F}}) = \sum_{j=1}^{l} \log p_\theta(y^j \mid \mathbf{x}^j)\,.$$

The maximization problem can be solved approximately by the SGD algorithm when the gradient of $\nabla_\theta L_{\mathrm{F}}(\theta; \mathscr{T}_{\mathrm{F}})$ is evaluated by the back-propagation.

B. WEAKLY ANNOTATED IMAGES

Second we consider weakly annotated images. In this case, the training set $\mathscr{T}_{\mathrm{W}} = \{(\mathbf{X}^j, y^j, c^j) \in \mathscr{X}^* \times \mathscr{Y} \times \mathscr{C} \mid j = 1, \ldots, m\}$ contains m triplets with the following meaning. The

Layer type	Configuration
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

TABLE I: 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'.

tensor $\mathbf{X}^j = (\mathbf{x}^j_1, \ldots, \mathbf{x}^j_{n^j}) \in \mathbb{R}^{100\times 100\times n_j}$ represents a bag of n^j sub-images which are cropped from the j-th database image around the bounding boxes found by a face detector and re-scaled to 100×100 pixels. The symbol $y^j \in \mathscr{Y}$ denotes the attributed label of a target person that should be captured in the j-th image. The symbol $c^j \in \mathscr{C} = \{1,\ldots,C\}$ denotes the identity label of the target person where C is the total number of identities in the database.

Given a weakly annotated image $(\mathbf{X}, y, c)$, we model the distribution of the label y conditioned on the identity c and the bag of detected face images $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ by

$$p_\theta(y \mid \mathbf{X}, c) = \sum_{\mathbf{h}\in\{0,1\}^n} p_\theta(y \mid \mathbf{X}, \mathbf{h})\, p_\theta(\mathbf{h} \mid \mathbf{X}, c)$$

where $\mathbf{h} = (h_1, \ldots, h_n) \in \{0,1\}^n$ is a vector of selector variables. The value $h_i = 1$ means that the i-th extracted image $\mathbf{x}_i$ contains the target person and $h_i = 0$ that it is someone else. Provided the selector variable $\mathbf{h}$ and the images $\mathbf{X}$ are given, the label distribution is simply

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

where $\mathbf{e}_i \in \{0,1\}^n$ is a vector of all zeros but the i-th component set to 1, $p_\theta(y \mid \mathbf{x}_i)$ models the attribute label distribution given the facial image and $p_\theta(y)$ is the label distribution in the database. The event $\mathbf{h} = \mathbf{e}_i$ means that just the i-th image out of $\mathbf{X}$ depicts the target person in which case the label y is distributed according to $p_\theta(y \mid \mathbf{x}_i)$. If no image is uniquely determined as the target person the label is distributed according to $p_\theta(y)$.

The distribution $p_\theta(\mathbf{h} \mid \mathbf{X}, c)$ models the appearance of the identity c. We assume that it is of the form

$$p_\theta(\mathbf{h} \mid \mathbf{X}, c) = \begin{cases} \tau & \text{if} \quad \mathbf{h} = \mathbf{0} \\ (1-\tau)\dfrac{\exp\langle \mathbf{w}_c, \phi(\mathbf{x}_i)\rangle}{\sum_{l=1}^n \exp\langle \mathbf{w}_c, \phi(\mathbf{x}_l)\rangle} & \text{if} \quad \mathbf{h} = \mathbf{e}_i \\ 0 & \text{otherwise} \end{cases}$$

The number $\tau \in [0,1]$ is the probability of the event $\mathbf{h} = \mathbf{0}$ when no images in the bag $\mathbf{X}$ depicts the target person, e.g. when the person is not visible in the scene or the detector fails. We assume that τ can be different for each database image but its average over the images associated to a single identity is equal a constant τ_0. That is, we assume

$$\frac{1}{|\mathscr{J}_c|}\sum_{j\in\mathscr{J}_c} \tau^j = \tau_0\,, \quad \tau^j \in [0,1]\,, j \in \{1,\ldots,m\}\,, \qquad (1)$$

where $\mathscr{J}_c = \{j \in \{1,\ldots,m\} \mid c^j = c\}$ are indices of database images assigned to identity c. In other words, τ_0 is our prior on the portion of database images for identity c where the detector failed to localize the identity. Unlike other parameters, τ_0 is not estimated from the data. We used $\tau_0 = 0.1$ in all experiments.

The probability that the i-th image from the bag $\mathbf{X}$ depicts the identity c is proportional to $\exp\langle \mathbf{w}_c, \phi(\mathbf{x}_i)\rangle$. The vector $\mathbf{w}_c \in \mathbb{R}^{d_I}$ is the template of the identity c and $\phi(\mathbf{x}_i) \in \mathbb{R}^{d_I}$ is the identity descriptor extracted from image $\mathbf{x}_i$. In our experiments, the descriptor is 4096-dimensional L_2-normalized output of the penultimate layer of the VGG-Face CNN [Parkhi et al., 2015]. In the course of EM we train only the identity templates $\mathbf{w}_c$ while the VGG-Face descriptor $\phi(\mathbf{x}_i) \in \mathbb{R}^{d_I}$ is fixed.

The conditional log-likelihood of θ given the weakly annotated training set $\mathscr{T}_\mathrm{W}$ reads

$$\begin{aligned} L_\mathrm{W}(\theta; \mathscr{T}_\mathrm{W}) &= \sum_{j=1}^m \log p_\theta(y^j \mid \mathbf{X}^j, c^j) \\ &= \sum_{j=1}^m \log \sum_{\mathbf{h}\in\{0,1\}^{n^j}} p_\theta(y^j \mid \mathbf{X}^j, \mathbf{h}) p_\theta(\mathbf{h} \mid \mathbf{X}^j, c^j)\,. \end{aligned}$$

III. ESTIMATING THE MODEL PARAMETERS BY EM ALGORITHM

We want to exploit both the fully annotated $\mathscr{T}_\mathrm{F}$ and weakly annotated examples $\mathscr{T}_\mathrm{W}$. The joint conditional log-likelihood reads

$$L(\theta; \mathscr{T}_\mathrm{F}, \mathscr{T}_\mathrm{W}) = L_\mathrm{F}(\theta; \mathscr{T}_\mathrm{F}) + L_\mathrm{W}(\theta; \mathscr{T}_\mathrm{W})\,,$$

where $\theta \in \mathbb{R}^d$ encapsulates all model parameters, namely, the convolution filters of the CNN defining the label distribution $p_\theta(y \mid \mathbf{x})$, prior distribution of the labels $p_\theta(y)$, the identity templates $\mathbf{w}_c, c \in \mathscr{C}$, and scalars τ^j, $j \in \{1,\ldots,m\}$, defining the probability that target identity was not among the faces detected in the j-th database image. We find the model parameters θ by solving

$$\theta^* = \operatorname*{argmax}_{\theta\in\mathbb{R}^d} L(\theta; \mathscr{T}_\mathrm{F}, \mathscr{T}_\mathrm{W}) \quad \text{subject to (1).} \qquad (2)$$

We solve the problem (2) approximately by an instance of the EM algorithm [Schlesinger, 1968], [Dempster et al., 1977].

The EM algorithm replaces the objective $L(\theta; \mathscr{T}_{\mathrm{F}}, \mathscr{T}_{\mathrm{W}})$ by the surrogate

$$F(\theta, \hat{p}) = \sum_{j=1}^{l} \log p_\theta(y^j \mid \mathbf{x}^j) p(\mathbf{x}^j) + \sum_{j=1}^{m} \sum_{\mathbf{h}} \hat{p}^j(\mathbf{h}) \log \Big(p_\theta(y^j \mid \mathbf{X}^j, \mathbf{h})\; p_\theta(\mathbf{h} \mid \mathbf{X}^j, c^j) \Big)$$

where $\hat{p} = \{\hat{p}^1, \ldots, \hat{p}^m\}$ is a collection of probability distributions over the hidden selector variables. Starting from an initial estimate θ_0, the EM performs the block coordinate ascent of $F(\theta, \hat{p})$:

EM algorithm solving the problem (2)

repreat
E-step: $\hat{p}_t = \operatorname{argmax}_{\hat{p}} F(\theta_{t-1}, \hat{p})$
M-step: $\theta_t = \operatorname{argmax}_{\theta} F(\theta, \hat{p}_t)$ s.t. (1)
until convergence

The surrogate function $F(\theta, \hat{p})$ is a tight lower bound of the log-likelihood $L(\theta; \mathscr{T}_{\mathrm{F}}, \mathscr{T}_{\mathrm{W}})$, namely, it holds that: i) $F(\theta, \hat{p}) \leq L(\theta; \mathscr{T}_{\mathrm{F}}, \mathscr{T}_{\mathrm{W}})$, $\forall \hat{p}$, $\forall \theta$, and ii) $\max_{\hat{p}} F(\theta, \hat{p}) = L(\theta; \mathscr{T}_{\mathrm{F}}, \mathscr{T}_{\mathrm{W}})$, $\forall \theta$. This implies that the EM algorithm monotonically increases the log-likelihood $L(\theta; \mathscr{T}_{\mathrm{F}}, \mathscr{T}_{\mathrm{W}})$. In our experiments we stopped the algorithm after 10 iteration when the improvement became negligable.

The optimization problems emerging in the E-step and the M-step are considerably simpler than the original problem (2). In particular, maximization in the E-step decomposes into m independent problems with closed form solution

$$\hat{p}_t^j(\mathbf{h}) = \frac{p_{\theta_{t-1}}(y^j \mid \mathbf{X}^j, \mathbf{h})\; p_{\theta_{t-1}}(\mathbf{h} \mid \mathbf{X}^j, c^j)}{\sum_{\mathbf{h}' \in \{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{n^j}\}} p_{\theta_{t-1}}(y^j \mid \mathbf{X}^j, \mathbf{h}')\; p_{\theta_{t-1}}(\mathbf{h}' \mid \mathbf{X}^j, c^j)}$$

for all $j \in \{1, \ldots, m\}$. To solve the maximization problem in the M-step, it is useful to rewrite the EM objective as

$$F(\theta, \hat{p}) = F_A(\theta, \hat{p}) + F_B(\theta, \hat{p}) + F_C(\theta, \hat{p}) + F_D(\theta, \hat{p})$$

where

$$F_A(\theta, \hat{p}) = \sum_{j=1}^{l} \log p_\theta(y^j \mid \mathbf{x}^j) + \sum_{j=1}^{m} \sum_{i=1}^{n^j} \hat{p}^j(\mathbf{h} = \mathbf{e}_i) \log p_\theta(y^j \mid \mathbf{x}_i^j)$$

$$F_B(\theta, \hat{p}) = \sum_{j=1}^{m} \hat{p}^j(\mathbf{h} = \mathbf{0}) \log p_\theta(y^j)$$

$$F_C(\theta, \hat{p}) = \sum_{j=1}^{m} \hat{p}^j(\mathbf{h} = \mathbf{0}) \log \tau^j + \sum_{j=1}^{m} \sum_{i=1}^{n^j} \hat{p}^j(\mathbf{h} = \mathbf{e}_i) \log(1 - \tau^j)$$

$$F_D(\theta, \hat{p}) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \hat{p}^j(\mathbf{h} = \mathbf{e}_i) \langle \mathbf{w}_c, \phi(\mathbf{x}_i^j) \rangle - \sum_{j=1}^{m} \sum_{i=1}^{n_j} \hat{p}^j(\mathbf{h} = \mathbf{e}_i) \log \sum_{l=1}^{n_j} \exp \langle \mathbf{w}_c, \phi(\mathbf{x}_l^j) \rangle$$

It seen that each function depends on a different set of parameters which allows to decompose the maximization problem into independent sub-problems whose solution is outlined below.

Sub-problem A. The maximization of $F_A(\theta, \hat{p}_t)$ w.r.t convolution filters contained in θ is equivalent to the supervised training of the CNN using the standard soft-max loss. We solve this problem approximately by performing 10 epochs of the SGD when the gradients of the CNN are computed by the back-propagation.

Sub-problem B. The maximization of $F_B(\theta, \hat{p}_t)$ w.r.t. distribution $p_\theta(y)$ contained in θ has analytic solution

$$p_{\theta_t}(y) = \frac{\sum_{j=1}^{m} [\![y^j = y]\!] \hat{p}_t^j(\mathbf{h} = \mathbf{0})}{\sum_{j=1}^{m} \hat{p}_t^j(\mathbf{h} = \mathbf{0})}, \quad y \in \mathscr{Y}.$$

Sub-problem C. The maximization of $F_C(\theta, \hat{p}_t)$ w.r.t. numbers τ^j contained in θ leads to C (number of identities in the database) independent convex problems

$$\max_{\tau \in \mathbb{R}^{m_c}} \sum_{j \in \mathscr{J}_c} \hat{p}_t^j(\mathbf{h} = 0) \log \tau^j + \sum_{j \in \mathscr{J}_c} \sum_{i=1}^{n^j} \hat{p}_t^j(\mathbf{h} = \mathbf{e}_i) \log(1 - \tau^j)$$

subject to (1). Solving the first order optimality conditions of the problem, i.e. the gradient of the problem's Lagrangian set to zero, leads to finding a root of an univariate function. We find the root by a binary search from which the optimal values of τ^j, $j \in \mathscr{J}_c$, are computed immediately.

Sub-problem D. The maximization of $F_D(\theta, \hat{p}_t)$ w.r.t. the identity descriptors $\mathbf{w}_c$, $c \in \mathscr{C}$, contained in θ leads to C independent convex unconstrained problems. We solve the problems by the BFGS algorithm.

IV. DATASETS

Two dataset corpora were used in the following experiments: the Clean dataset and the IMDB dataset.

a) Clean dataset: The dataset has full ground-truth annotation of age and gender associated to every face. The dataset consists of 87,485 images in total. The training (78%), validation (4%), and test (18%) split was kept fixed. The training and test data are composed from 70% of PubFig [Kumar et al., 2009] and 30% of LFW datasets [Huang et al., 2007], while the test data are composed from additional datasets, in summary Pubfig 55%, ChaLearnAge 22%, LFW 17% and FG-NET 22%. The age range is from 16 to 75 and there is 43% of female subjects.

b) IMDB dataset: The dataset collected by [Rothe et al., 2015] 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 photo of a photo (George Washington 300 years, J. F. Kennedy 90 years, Jack London 134 years), or due to name coincidence

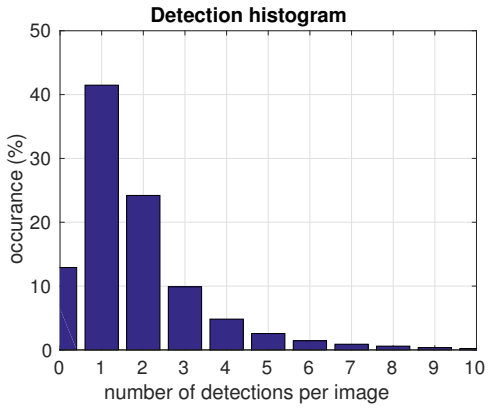


Fig. 2: Histogram of number of detected faces per image in the IMDB dataset.

(Kathryn Boyd 110 years). We finally discarded all images having age out of the range [16, 75] or invalid gender.

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[1] 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. 2.

V. BASELINES

In this section we describe 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.

A. Baseline 1: Dominant face detection

In the first approach to the selection problem we follow [Rothe et al., 2015]. 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 the detection score. The minimum detection score to output a bounding box is 30. The second condition adds about 7k images. In total, the final *baseline-1* dataset contains 194,540 images.

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, 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.

B. Baseline 2: Median identity from single detections

We propose another selection strategy which exploits the annotation of the person identity. For all detections, an identity descriptor $\phi(\mathbf{x})$ is computed. The descriptor is 4096-dimensional L_2-normalized output of the penultimate layer of the VGG-Face CNN [Parkhi et al., 2015]. 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. Altogether, the *baseline-2* dataset contains 313,520 images.

[1]Eyedea Recognition, Ltd. www.eyedea.cz

VI. EXPERIMENTS

A. 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.

The network was trained by using various baseline training sets: clean, *baseline-1*, *baseline-2*, clean + *baseline-1*, clean + *baseline-2* which are fully labelled, and by using the proposed EM training on the entire weakly annotated IMDB dataset (EM-CNN). The training is repeated three times always starting from initial random weights of the network. Besides the error statistics, a standard deviation is provided.

The results are summarized in Tab. II. It is seen that training from the small clean dataset has the worst error statistics. Training from *baseline-1* has better results and even better results are achieved by training from *baseline-2*. Unifying the *baseline-1* and *baseline-2* with the small clean dataset does not have a significant impact on the results. The proposed EM-CNN training outperforms all baselines in MAE and CS5, and is very similar in gerr to *baseline-2*. However the difference is not very significant. The results indicate 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.

Nevertheless, 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. This dataset is created artificially by erasing all single detections from IMDB dataset, dropping about 186k images. We select a subset *baseline-2* (woSingle) the same as *baseline-2* 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 Tab. II it is seen that proposed EM-CNN (woSingle) training outperforms the other baseline approaches *baseline-2* (woSingle) and clean + *baseline-2* (woSingle) by a significant margin. Note that results of the EM-CNN (woSingle) stayed almost the same compared to the case with the full

method	number of training samples annotated		MAE	CS5 [%]	gerr [%]
	fully	weakly			
clean	67,832	0	6.41 ± 0.20	51.67 ± 1.06	4.88 ± 0.44
baseline-1	0	187,211	5.80 ± 0.39	57.45 ± 2.81	3.35 ± 0.11
baseline-2	0	301,871	5.23 ± 0.05	61.88 ± 0.26	$\mathbf{2.73} \pm 0.07$
clean + *baseline-1*	67,832	187,211	5.40 ± 0.07	59.01 ± 0.52	3.47 ± 0.59
clean + *baseline-2*	67,832	301,871	5.35 ± 0.28	60.28 ± 2.51	2.79 ± 0.16
EM-CNN	67,832	859,198	$\mathbf{5.06} \pm 0.01$	$\mathbf{62.48} \pm 0.31$	2.74 ± 0.06
baseline-2 (woSingle)	0	103,691	6.68 ± 0.24	50.43 ± 1.58	4.22 ± 0.26
clean + *baseline-2* (woSingle)	67,832	103,691	5.97 ± 0.49	54.37 ± 4.14	3.59 ± 0.65
EM-CNN (woSingle)	67,832	751,798	$\mathbf{5.06} \pm 0.71$	$\mathbf{62.65} \pm 0.22$	$\mathbf{2.67} \pm 0.07$

TABLE II: The results. Besides the error statistics of the methods (MAE, CS5, gerr), the table shows the number of training samples of face images taken from the fully and weakly annotated datasets. Note that a small portion of datasets presented in Sec. IV was used for validation.

IMDB dataset (EM-CNN) in spite of dropping 186k single detection images.

The results of age estimation are further illustrated in Fig. 3. The plots show the mean average error as a function of age category. The plots are shown for both the entire IMDB dataset (a) and for the case when single detections are not considered (b). It is seen that for the first case, the EM-CNN is the best except for ages below 30. While for the latter case, the error of the proposed EM-CNN (woSingle) is always lower than other baselines over all age categories.

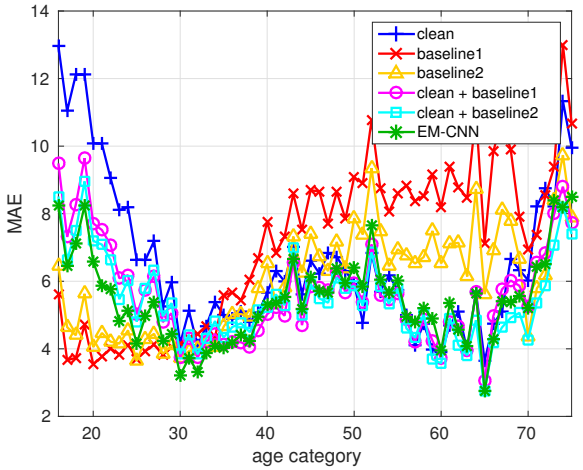
B. Sensitivity on the correctness of the training data

To assess a sensitivity of the final CNN classifier on the errors in the training labels, the following experiment was performed. The clean dataset $\mathscr{T}_{\text{clean}}$ was extended with an increasingly larger portion of images from the IMDB dataset with labels generated randomly from uniform distribution over the classes. We call the extra set here the outlier dataset, $\mathscr{T}_{\text{outlier}}$. The network was trained from the union of the clean and outlier sets always from scratch, i.e. with random initial weights. The ratio of outliers in the composed training set is $r = \frac{|\mathscr{T}_{\text{clean}}|}{|\mathscr{T}_{\text{clean}} \cup \mathscr{T}_{\text{outlier}}|}$.

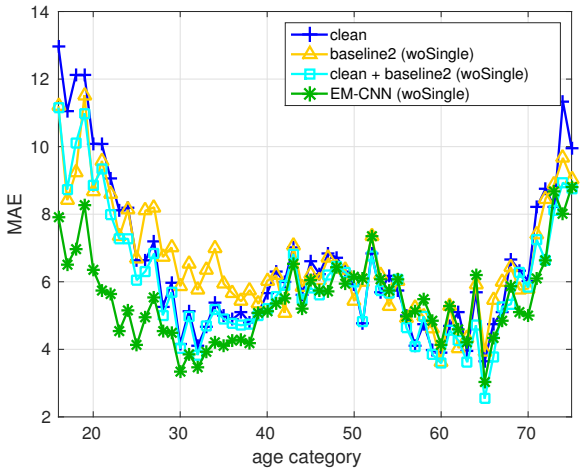
Results are shown in Fig. 4 as plots of gender error and age mean absolute error as a function of outlier ratio r. We can see that both error statistics deteriorates with increasing ratio of outliers. Nevertheless, the drop of recognition accuracy is surprisingly slow considering the number of outliers. In the extreme case, there is more outliers than correct samples and still the training process does not fail completely. This little sensitivity can probably be explained by the noise uniformity causing that erroneous sub-gradients are averaged out throughout the SGD training.

C. Purity of the training dataset

Beside the overall accuracy of the trained network, a quality of the training set was evaluated by measuring correctness of the assignment of the target person to detections on IMDB images. We have manually annotated a set of 960 images from the IMDB dataset. Two subsets of equal size were annotated: the images with single detections only, and the images with multiple detections. Each of the subsets was randomly sampled such that all age and gender categories



(a) All IMDB images.



(b) IMDB images without single detections.

Fig. 3: Mean Average Error of the age estimate per age categories. Results for datasets derived from: the complete IMDB dataset (a), and the subset of IMDB dataset without using images having single detections only (b).

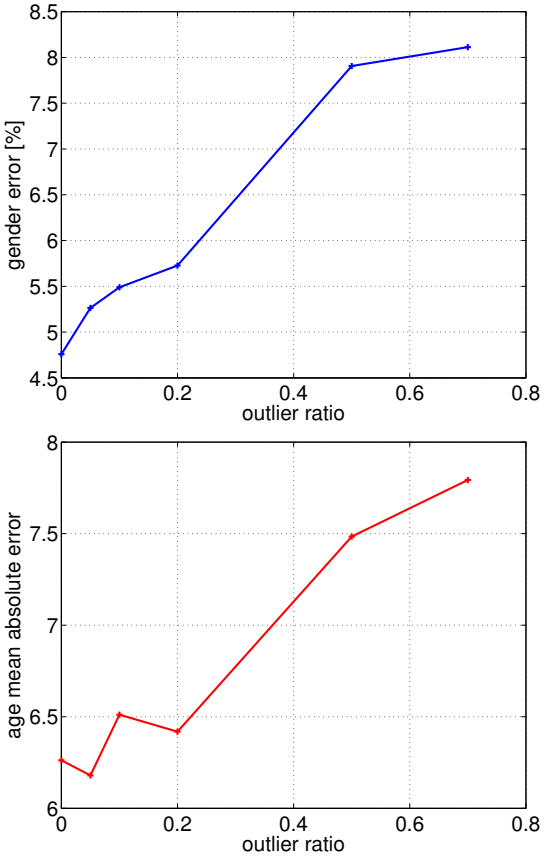


Fig. 4: Training with outliers. The basic training set (clean) was successively extended with images with randomly generated labels, the outliers. Accuracy of the trained network deteriorates but surprisingly slowly.

are uniformly present, i.e. two samples from each category of gender and age $\{F,M\} \times \{16,\ldots,75\}$. 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 not so easy for human. The assignment is not always unambiguous, especially for low resolution and non-frontal views.

Having set of N images with the ground-truth assignments, three errors were measured for each of the training sets: False Negatives (FN) as number of images that contain the target person and none of the detections appeared in the training set, False Positives (FP) as number of images that does not contain the target person and any of the detection appears in the training set, and Mismatches (MI) as number of images where a wrong detection is selected instead of the correct one. The overall assignment error is $\text{err} = \frac{FN+FP+MI}{N}$.

To evaluate the EM approach, that outputs probability distribution over assignments $\hat{p}^j$, we selected the assignment with the maximum probability including the case where the target person not present is the most probable.

method	FN	FP	MI	err [%]
baseline-1	**0**	103	0	21.46
baseline-2	13	**21**	0	**7.08**
EM-CNN	24	53	0	16.04

(a) Single detections only (480 images).

method	FN	FP	MI	err [%]
baseline-1	421	2	**4**	88.96
baseline-2	55	**1**	14	**14.58**
EM-CNN	**29**	36	53	24.58
baseline-2 (woSingle)	197	**15**	52	55.00
EM-CNN (woSingle)	**22**	47	**28**	**20.21**

(b) Multiple detections (480 images).

TABLE III: Purity of a training database.

The results are presented for single detection images and multiple detection images in Tab. IIIa and Tab. IIIb respectively. We can see that the first selection strategy *baseline-1* works rather well for single detections, although high FP is caused by no mechanism to suppress wrong single detections. For multiple detections, *baseline-1* is clearly sub-optimal, since only a small fraction of such images is selected resulting in high FN. The second selection strategy *baseline-2* works much better having the lowest assignment error for both single and multiple detections among all strategies. The EM approach is slightly worse in the assignment. The main reason we see is that the method decides the assignment based on the age and gender beside the face similarity, and thus for instance tend to assign wrong detection if the age and gender matches.

The model of the EM approach does not take any prior information from the single detections, which is apparently a very strong cue in this particular IMDB dataset. However, on a dataset where this prior does not exist, the general model works fine. The EM-CNN (woSingle), which is using only the subset of IMDB images without single detections, outperforms the *baseline-2* (woSingle). The simple heuristic based on a distance to a single etalon *baseline-2* (woSingle) is not working in this case.

VII. CONCLUSIONS

In this paper, we have addressed a problem of learning CNNs for face recognition from weakly annotated images. A weakly annotated image is assigned a pair of a attribute label and an identity label corresponding to a single person in the image. It is unknown which face extracted from the image corresponds to the annotation. It is further assumed that each identity is associated with more than one image in the database.

We have proposed a novel heuristic for assigning the annotations to faces. The heuristic exploits images with a single face detection to build an identity model which is subsequently used to select the correct faces. The proposed heuristic creates a cleaner annotation than the so far use existing heuristic ignoring the identity. The cleaner annotation allows to train a CNN for age and gender estimation with

a significantly higher accuracy than the baseline proposed by [Rothe et al., 2015].

The main contribution is a principled approach which formulates learning from weakly annotated images as a maximum likelihood estimation of a parametric distribution describing the data. The ML problem can be solved by an instance of the EM algorithm which in its inner loop learns a CNN for given face recognition problem. Experiments on age and gender estimation problem show that the proposed EM-CNN algorithm consistently outperforms the heuristic approaches. Unlike the heuristic methods, the EM-CNN does not require images with a single detection. Moreover, the EM-CNN has a single hyper-parameter corresponding to the portion of images in which the target person was not detected. In contrast, the heuristic methods require a set of well tuned thresholds.

REFERENCES

[Andrews et al., 2002] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *Proc. of Neural Information Processing Systems*.

[Antipov et al., 2016] Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (2016). Apparent age estimation from face images combining general and children-specialized deep learning models. In *CVPR workshop, Looking at People Challenge*.

[Antoniuk et al., 2016] Antoniuk, K., Franc, V., and Hlavac, V. (2016). V-shaped interval insensitive loss for ordinal classification. *Machine Learning*.

[Chang et al., 2011] Chang, K.-Y., Chen, C.-S., and Hung, Y.-P. (2011). Ordinal hyperplane ranker with cost sensitivities for age estimation. In *CVPR*.

[Chen et al., 2015] Chen, C., Patel, V., and Chellappa, R. (2015). Matrix completion for resolving label ambiguity. In *IEEE International Conference on Computer Vision (ICCV)*.

[Cour et al., 2011] Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, 12:1225–1261.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38.

[Escalera et al., 2015] Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H. J., and Guyon, I. (2015). Chalearn 2015 apparent age and cultural event recognition: datasets and results. In *ICCV, ChaLearn Looking at People workshop*.

[Geng et al., 2010] Geng, X., Smith-Miles, K., and Zhou, Z. (2010). Facial age estimation by learning from label distributions. In *Proc. of Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[Geng et al., 2007] Geng, X., Zhou, Z., and Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Analysis and Machine Learning*, 29(12):2234–2240.

[Han et al., 2013] Han, H., Otto, C., and Jain, A. K. (2013). Age estimation from face images: Human vs. machine performance. In *International Conference on Biometrics (ICB)*.

[Huang et al., 2007] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[Jim and Ghahramani, 2002] Jim, R. and Ghahramani, Z. (2002). Learning with multiple labels. In *Proc. of NIPS*.

[Kumar et al., 2009] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*.

[Lanitis et al., 2002] Lanitis, A., Taylor, C., and Cootes, T. (2002). Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Analysis and Machine Learning*, 24(4):442–455.

[Panis et al., 2016] Panis, G., Lanitis, A., Tsapatsoulis, N., and Cootes, T. (2016). Overview of research on facial ageing using the fg-net ageing database. *IET Biometrics*, 5(2).

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

[Ricanek and Tesafaye, 2006] Ricanek, K. J. and Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *IEEE 7th International Conference on Automatic Face and Gesture Recognition*, pages 341–345, Southampton, UK.

[Rothe et al., 2015] Rothe, R., Timofte, R., and Gool, L. V. (2015). Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[Schlesinger, 1968] Schlesinger, M. (1968). A connection between learning and self-learning in the pattern recognition (in Russian). *Kibernetika*, 2:81–88.

[Yan et al., 2008] Yan, S., Wang, H., Tang, X., Liu, J., and Huang, T. (2008). Regression from uncertain labels and its applications to soft biometrics. *IEEE Transactions on infromation forensics and security*, 3(4):698–708.