Structured Output SVM Prediction of Apparent Age, Gender and Smile From Deep Features

Michal Uřičář
CMP, Dept. of Cybernetics
FEE, CTU in Prague
uricamic@cmp.felk.cvut.cz

Radu Timofte
Computer Vision Lab
D-ITET, ETH Zurich
radu.timofte@vision.ee.ethz.ch

Rasmus Rothe
Computer Vision Lab
D-ITET, ETH Zurich
rrothe@vision.ee.ethz.ch

Jiří Matas
CMP, Dept. of Cybernetics
FEE, CTU in Prague
matas@cmp.felk.cvut.cz

Luc Van Gool
PSI, ESAT, KU Leuven
CVL, D-ITET, ETH Zurich
vangool@vision.ee.ethz.ch

Abstract

We propose structured output SVM for predicting the apparent age as well as gender and smile from a single face image represented by deep features. We pose the problem of apparent age estimation as an instance of the multi-class structured output SVM classifier followed by a softmax expected value refinement. The gender and smile predictions are treated as binary classification problems. The proposed solution first detects the face in the image and then extracts deep features from the cropped image around the detected face. We use a convolutional neural network with VGG-16 architecture [25] for learning deep features. The network is pretrained on the ImageNet [24] database and then fine-tuned on IMDB-WIKI [21] and ChaLearn 2015 LAP datasets [8]. We validate our methods on the ChaLearn 2016 LAP dataset [9]. Our structured output SVMs are trained solely on ChaLearn 2016 LAP data. We achieve excellent results for both apparent age prediction and gender and smile classification.

1. Introduction

While the problem of the physical (*i.e.* the biological, real) age estimation has been covered by numerous studies [1, 4, 7, 12, 22], the estimation of the apparent age is still in its beginnings and just recently it is getting the well deserved attention. There are various applications of apparent age estimation systems, such as medical diagnosis (premature aging), forensics, plastic surgery, to name a few. The biggest barrier in apparent age research is the lack of larger annotated datasets, which should also cover the uncertainty of the human annotators.

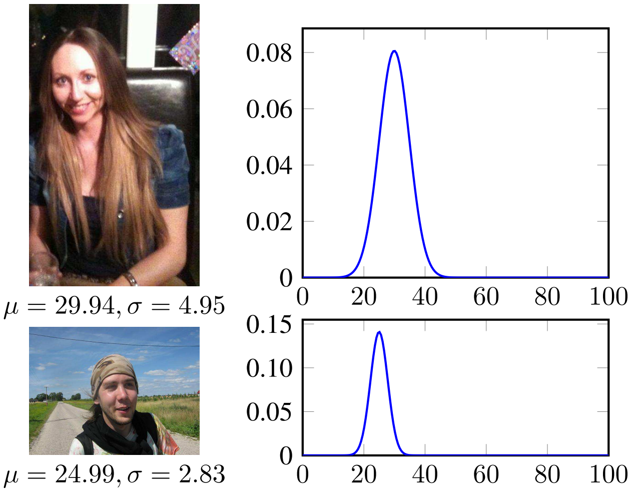


Figure 1. Exemplary images from the ChaLearn LAP 2016 dataset [9] and their annotations. The graphs show the uncertainty of the annotators, measured by the standard deviation.

ChaLearn Looking At People (LAP) 2015 challenge [8] proposed a special track on apparent age estimation and provided the largest public dataset of annotated face images for apparent age estimation at that time. ChaLearn LAP 2016 [9] extends the LAP 2015 dataset with new images and annotations. Figure 1 depicts exemplary images from the newly released dataset. The apparent age estimation system should detect faces in the image and output the prediction. Note that the images are taken in the wild (uncontrolled environment).

In this work, we build on the approach of Rothe *et al.* [21], winner of the ChaLearn LAP 2015 challenge on apparent age estimation [8], and explore the potential of the Structured Output Support Vector Machines (SO-SVM) [28] in combination with deep features extracted from face images. The recent advances in computer vision applications of deep learning [5, 15, 24] and deep features [6, 11, 13, 17], but also the deep learning winner solution of ChaLearn LAP 2015 give us the motivation to use deep features, while the theoretical derivation and its robustness vouch for SO-SVM.

We use the VGG-16 architecture for our convolutional neural network (CNN), which was pre-trained on the ImageNet [24] for image classification, modified for the apparent age estimation and fine-tuned on the IMDB-WIKI dataset [21] (annotated with the physical age) and ChaLearn LAP 2015 [8] dataset (annotated with the apparent age). This network showed excellent results in the ChaLearn 2015 challenge, outperforming all the competitors [21]. In this work, we show that the results can be significantly improved by taking the deep features and formulating a SO-SVM multi-class classifier on top of it. The main benefits of SO-SVM algorithm are i) usage of an arbitrary loss function, which can therefore be identical to the testing error measure ii) existence of the ε-precise solvers. Our main contributions are as follows

1. a novel approach to apparent age, gender and smile prediction using SO-SVM and deep features.
2. a robust system with excellent prediction performance validated on the latest ChaLearn LAP challenges.
3. experimental results verifying that the SO-SVM in combination with learned deep features is significantly better than a direct deeply learned prediction.

1.1. Related work

The most relevant works in apparent age estimation and related to ours are the ones published as result of the ChaLearn LAP 2015 challenge [8]. In the next we briefly review the top works starting with the winning solution which constitutes the starting point for our SO-SVM solution and then briefly review related works on gender and smile estimation in the wild.

1.1.1 Apparent age estimation

Rothe *et al*. [21], 1^{st} place at ChaLearn LAP 2015 challenge, propose the Deep Expectation (DEX) method. DEX employs a robust face detection and alignment based on the detector of Mathias *et al.* [20] and an ensemble prediction with 20 CNNs applied on the cropped face. Each network has the VGG-16 architecture [25] and is pre-trained on ImageNet and then fine-tuned on face images from IMDB-WIKI with the physical age and ChaLearn LAP 2015 with the apparent age annotations. The networks are trained to predict with 101 output neurons, each neuron corresponds to a discretized age from the interval $[0, \ldots, 100]$. The final prediction is the expected value of the softmax-normalized output of the last layer, averaged over the ensemble of 20 networks.

Liu *et al*. [18], 2^{nd} place, also use an ensemble (of 10) of large scale deep CNNs based on GoogleNet architecture [26]. CASIA-WebFace database [31] and an large outside age dataset are used for training. Two kind of losses are used— Euclidean loss for single dimension age encoding and a cross-entropy loss of label distribution learning based age encoding. A face landmark detection is used for face alignment and a system prediction based on three cascaded CNNs for face classification, real age, and apparent age.

Zhu *et al*. [32], 3^{rd} place, combines face and face landmark detection with a GoogleNet deep architecture trained on $240,000$ public face images with physical age annotation consequently fine-tuned on augmented ChaLearn LAP 2015 dataset. Age grouping is further applied such that each face is classified into one of 10 age groups. Within each group random forest and support vector regression are used to train the age estimator. Final prediction is formed by score level fusion of individual classifiers.

1.1.2 Gender and smile prediction

As is the case for apparent age estimation, the recent years showed tremendous advances in the prediction of facial attributes such as smile or gender fueled by the newly introduced large datasets with annotations and by new methods or deep learning.

Recently, Azarmehr *et al*. [2] presented a complete framework for the video-based age and gender classification using non-linear SVMs with Radial Basis Function (RBF) kernel. Also, Liu *et al.* [19] introduced the CelebA database containing more than $200,000$ images with 40 annotated facial attributes, as well as a deep learning framework for attribute prediction in the wild. Li *et al.* [16] propose a binary code learning framework for facial attributes prediction and release a database called CFW 60K.

The rest of the paper is organized as follows. Section 2 introduces our method. Section 3 describes the experimental setup and the achieved results and discusses them. Finally, Section 4 concludes the paper.

2. Proposed method

The pipeline of our method is depicted in Figure 2. It exploits the deep features from DEX [21], the winner of the previous ChaLearn challenge on apparent age estimation [8], and learns an ensemble of linear SO-SVM classifiers with the desired loss function, *i.e.* an ϵ-error (12). Since we use a binary SVM classifier for the gender and smile prediction, we will describe just the experiments we made in Section 3.

In the next sections, we describe the deep features and final classifier in detail.

2.1. Deep features

We adopt the deep network from DEX [21]. Thus, we use the VGG-16 architecture [25] pre-trained on the ImageNet dataset [23] subsequently fine-tuned on the IMDB-WIKI dataset [21] and the ChaLearn LAP dataset [8].

In contrast to the DEX method, we use the deep network solely as a feature extractor. We have experimented with several different options for the features, like to use the fully connected layers (with or without ReLU), their combination or the convolutional layers with dimensionality reduction by PCA. The best results for the apparent age estimation were achieved with the last fully connected layer (fc7), without ReLU. For the gender and smile prediction, we got the best results using the last but one fully connected layer (fc6).

The features are extracted from the fixed size image, which is formed as follows. Firstly, the faces in the input image are detected by the off-the-shelf detector of Mathias *et al.* [20]. Since the faces in ChaLearn LAP datasets are in unconstrained poses, we rotate the input image in the interval of $[-60°, 60°]$ in $5°$ steps and also by $-90°, 90°$ and $180°$. The face box with the strongest detection score is taken, together with the rotation angle. In the rare case that no face is detected, we take the entire image for further processing. Secondly, the face box size is enlarged by 40% in both width and height and the face image is cropped. The resulting image is eventually squeezed to 256×256 pixels and used as an input to the deep convolutional network for the features extraction.

2.2. Structured Output SVM prediction

The age estimation can be posed as a multi-class classification task, where the classes $y \in \mathcal{Y}$ correspond to age discretized by years, *i.e.* $\mathcal{Y} = \{0, 1, \ldots, 100\}$. Given the input image $\boldsymbol{x} \in \mathcal{X}$, we are interested in the best performing classifier $h\colon \mathcal{X} \to \mathcal{Y}$, which minimizes the ϵ-error (12).

Following the SO-SVM framework [28], let us define a scoring function $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as a linear function of the parameters $\boldsymbol{w}$ to be learned from m fully annotated training examples $\mathcal{T} = \{(\boldsymbol{x}^i, y^i, \sigma_y^i)\}_{i=1}^m$, where σ denotes the standard deviation of the age label among annotators, and the features map $\boldsymbol{\Psi}(\boldsymbol{x}, y) \in \mathbb{R}^n$:

$$f(\boldsymbol{x}, y) = \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, y) \rangle \ . \tag{1}$$

The classifier then decides based on where the scoring function attends its maximum

$$h(\boldsymbol{x}; \boldsymbol{w}) = \arg\max_{y \in \mathcal{Y}} f(\boldsymbol{x}, y) \ . \tag{2}$$

We define the $\boldsymbol{\Psi}(\boldsymbol{x}, y) \in \mathbb{R}^n$ as follows

$$\boldsymbol{\Psi}(\boldsymbol{x}, y) = (0, \ldots, 0, \boldsymbol{x}, 0, \ldots, 0)^\top \ , \tag{3}$$

that is, we stack the feature vector $\boldsymbol{x}$ into position y of the zero vector, which has the dimensionality equal to $n = |\mathcal{Y}| \cdot \dim(\boldsymbol{x}) = 101 \cdot 4,096 = 413,696$.

The SO-SVM algorithm translates the problem of learning the classifier parameters $\boldsymbol{w}$ into the following convex task

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} F(\boldsymbol{w}) := \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} r_i(\boldsymbol{w}) \right] \ , \tag{4}$$

where $r_i(\boldsymbol{w})$ is a loss incurred by the classifier on the i-th training example $(\boldsymbol{x}^i, y^i, \sigma_y^i)$ and $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ is a quadratic regularizer introduced to prevent the over-fitting. The optimal value of the regularization constant λ is to be found in the model selection on the independent validation set. The loss $r_i(\boldsymbol{w})$ is the margin-rescaling convex proxy (*c.f.* [28]) of the true loss $\Delta(y, \sigma_y', y')$ defined as follows:

$$\begin{aligned} r_i(\boldsymbol{w}) = \max_{y \in \mathcal{Y}} \Big[\Delta(y, \sigma_y^i, y^i) \quad &+ \quad \big\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, y) \big\rangle \\ &- \quad \big\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, y^i) \big\rangle \Big] \ . \end{aligned} \tag{5}$$

The true loss $\Delta(y, \sigma_y', y')$ is set to be the ϵ-error (12). Note that the evaluation of the proxy loss $r_i(\boldsymbol{w})$ is equivalent to running the classifier with the scoring function augmented by the values of the true loss $\Delta(y, \sigma_y', y')$, and that the ϵ-error fulfills the requirements of the loss function posed by SO-SVM framework, because $\Delta(y, \sigma_y', y') = 0 \iff y = y'$.

We solve (4) approximately by the Bundle Methods for Regularized Risk Minimization (BMRM) algorithm [27], which is outlined in Algorithm 1. The core idea is to approximate the original hard problem (4) by its reduced problem

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} F_t(\boldsymbol{w}) := \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + r_t(\boldsymbol{w}) \right] \ , \tag{6}$$

where the objective $F_t(\boldsymbol{w})$ is constructed as a cutting plane model of the original risk term $r(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^m r_i(\boldsymbol{w})$

$$r_t(\boldsymbol{w}) = \max_{i=1,\ldots,t} \left[r(\boldsymbol{w}_i) + \langle \boldsymbol{r}'(\boldsymbol{w}_i), \boldsymbol{w} - \boldsymbol{w}_i \rangle \right] \ , \tag{7}$$

$\boldsymbol{r}'(\boldsymbol{w}) \in \mathbb{R}^n$ denotes the sub-gradient of $r(\boldsymbol{w})$ evaluated at point $\boldsymbol{w}_i \in \mathbb{R}^n$.

The BMRM algorithm starts from the initial guess $\boldsymbol{w}_0 = \boldsymbol{0}$ and in iteration t computes $\boldsymbol{w}_t$ by solving the reduced problem (6), by adding a cutting plane computed at the intermediate solution $\boldsymbol{w}_t$ to the cutting plane model (7). This leads to a progressively tighter approximation of $F(\boldsymbol{w})$.

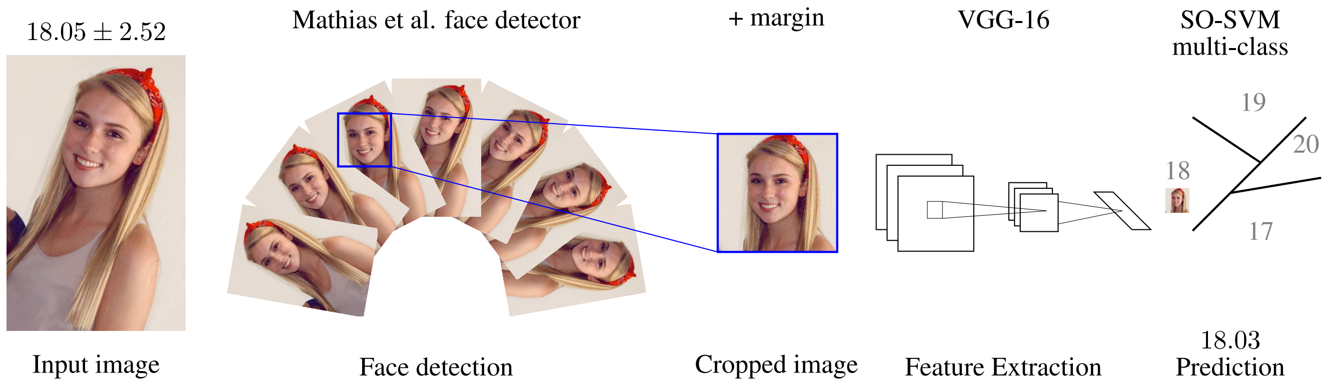


Figure 2. Proposed pipeline for the apparent age estimation.

Algorithm 1 BMRM algorithm

```
Require: $\varepsilon$, first order oracle evaluating $r(\boldsymbol{w})$ and $\boldsymbol{r}'(\boldsymbol{w})$
Initialization: $\boldsymbol{w} \leftarrow \mathbf{0}, t \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    Call oracle to compute $r(\boldsymbol{w}_t)$ and $\boldsymbol{r}'(\boldsymbol{w}_t)$
    Update the cutting plane model $r_t(\boldsymbol{w}_t)$
    Solve the reduced problem (6)
until $F(\boldsymbol{w}_t) - F_t(\boldsymbol{w}_t) \leq \varepsilon$
```

The BMRM was proven [27] to converge to an ε-precise solution (*i.e.* satisfying $F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$) in $\mathcal{O}(\frac{1}{\varepsilon})$ iterations for arbitrary $\varepsilon > 0$.

The BMRM algorithm requires a first order oracle which, for a given query $\boldsymbol{w}_t$, evaluates $r(\boldsymbol{w}_t)$ and the sub-gradient $\boldsymbol{r}'(\boldsymbol{w}_t) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{r}'_i(\boldsymbol{w}_t)$. The components of the sub-gradient $\boldsymbol{r}'_i(\boldsymbol{w}_t)$ are computed by the Danskin's theorem [3] as:

$$\boldsymbol{r}'_i(\boldsymbol{w}_t) = \boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{y}) - \boldsymbol{\Psi}(\boldsymbol{x}^i, y^i) \,, \tag{8}$$

where

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \left[\Delta(y, \sigma_y^i, y^i) + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{y}) \rangle \right] . \tag{9}$$

Since the LAP apparent age annotations are in floating point precision, we further investigated the possibility of giving real-valued prediction instead of the discretized y. Inspired by the DEX method [21], we compute the softmax assignment y_s based on the values of the scoring function $f(\boldsymbol{x}, y)$

$$s_i = \frac{e^{f(\boldsymbol{x}, y_i)}}{\sum e^{f(\boldsymbol{x}, y_i)}}, \forall i \in \mathcal{Y}, \quad y_s = \mathrm{E}[y] = \sum_{i=0}^{100} y_i s_i \,, \tag{10}$$

where s_i represent the softmax output probabilities and $y_i \in \mathcal{Y}$ are the discrete years.

3. Experiments

In this section we first briefly describe the datasets and the evaluation protocol from our experiments. Then we provide the implementation details for our method and discuss the results.

3.1. Datasets and evaluation protocol

3.1.1 Apparent age estimation

In our experimental validation we use the ChaLearn LAP 2016 dataset [9] as provided by the challenge organizers. Note that we use only the ChaLearn LAP 2016 dataset for training our multi-class SO-SVM classifier. This dataset has $5,613$ face images ($4,113$ for training and $1,500$ for validation) with age annotations and $1,978$ face images for testing (annotations are not publicly available). The annotation consists of the mean value of the independent annotator guess of the apparent age and the standard deviation σ of this guess across the annotators.

The histograms of the apparent age of both training and validation parts of the ChaLearn LAP 2016 database [9] are depicted in Figure 3. Note that the age distributions for both parts are approximately the same, covering the best the 20–40 years interval. Another significant peak is visible for the interval 0–15 and around 50, while the years higher than 90 are left completely uncovered.

We evaluate the results by two different statistics. The first is the mean absolute error (MAE), computed as the average of absolute errors between the predictions and the ground truth age annotations:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m} |y_i - y^i| \,, \tag{11}$$

where m is the number of predictions. The second is the ϵ-error used by the ChaLearn LAP challenge:

$$\epsilon = 1 - e^{-\frac{(y - y^i)^2}{2\sigma^2}} \,. \tag{12}$$

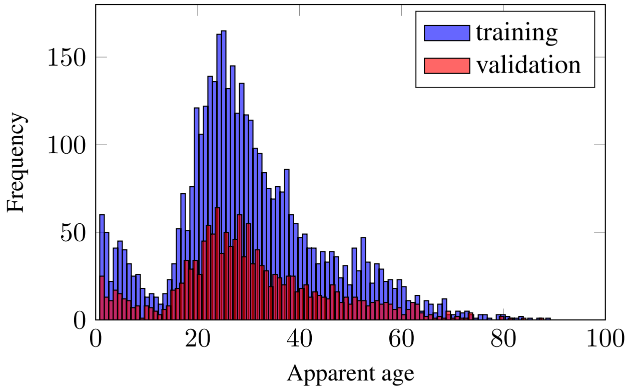


Figure 3. Histograms of apparent age of the training and validation sets of the ChaLearn LAP 2016 dataset [9].

For a set of m images, we report the average of eq. (12) acquired for individual images. Note that the maximum value of the ϵ-error is equal to 1, which represents the worst result and the minimum is equal to 0 representing the best match. Even though MAE does not take into account the uncertainty of the ground truth age, it is still a useful measure of the prediction accuracy.

3.1.2 Gender and smile prediction

We use the ChaLearn LAP 2016 dataset [9] for gender and smile detection. The dataset consists of $9,257$ face images in total ($6,171$ for training and $3,086$ for validation) with gender (0-male / 1-female) and smile (1-yes / 0-no) annotations. The testing set was kept secret (both the images and the annotations).

We evaluate the classifiers independently for the development purposes, by simply calculating the average mismatch of the predicted and ground truth labels. The final evaluation is done jointly, *i.e.* the mean square error between the prediction and the ground truth is calculated. The error ranges between 0 (both gender and smile prediction matches the ground truth) and 1 (both gender and smile are misclassified). In case a face was not detected, the error is set to 1.

3.2. Implementation details

3.2.1 Apparent age estimation

We train 10 different versions of the described multi-class SO-SVM classifier for the apparent age estimation on 10 different splits of the new ChaLearn LAP 2016 dataset [9] (both training and validation sets from LAP are included). The final prediction is the average of the ensemble predictions of these 10 classifiers.

Both the training of the proposed multi-class SO-SVM classifier and the pipeline for prediction of the apparent age are coded in MATLAB. We use the Caffe framework [14] for the CNNs and deep features extraction. For the face detection, we use the detector from [20].

The features extraction takes around 200 ms per image (on a GeForce GTX Titan Black graphics card). The final classifier (ensemble of 10 multi-class SO-SVM classifiers enhanced by softmax expectation) takes approximately 0.5 ms. The slowest part is the face detection, which takes 280–420 s per image, however it can be easily parallelized. The training of the whole ensemble of SO-SVM classifiers took 1 day for the whole model selection (*i.e.* finding the optimal value of the regularization parameter λ in (6)). We used only the ChaLearn LAP 2016 [9] dataset for training. The source codes of both training scripts and final proposed classifier are publicly available at:

http://cmp.felk.cvut.cz/~uricamic/LAP2016/

3.2.2 Gender and smile prediction

We train a single classifier per each prediction task employing the OCAS algorithm [10]. We use fc6 deep features extracted from the VGG-16 network pretrained on ImageNet and fine-tuned for gender classification on the IMDB-WIKI database. The input image for the CNN is constructed similarly as in the apparent age estimation case.

In the development phase, we used the commercial face detector[1], implementing the Waldboost detector [30]. This face detector is much faster than [20] (approx. 4 seconds for all image rotations, compared to the 280–420 seconds), however, the scale and position of the detected face box are not so stable. Therefore, we correct the face box position and scale based on the facial landmarks. We use the CLandmark [29] facial landmark detector with 2 different models— the first one is the multi-view detector, which besides of the facial landmarks provides the discretized yaw head-pose information, the second one is the coarse-to-fine detector, which detects 68 landmarks for near-frontal yaw poses. We use the detected landmarks to construct the corrected face box (we use the 12 eye landmarks from the 68 landmarks set detected for near-frontal poses for correction of the in-plane rotation and face box size, and eye landmarks and vertical face size for non-frontal poses), which is then used to form the input image supplied to the CNN feature extractor.

We observed that the quality of the constructed input image using this approach is practically the same compared to the approach used for the apparent age estimation. However, the Waldboost detector in combination with landmark detection removes the biggest computational bottleneck and is therefore more suitable for real-time applications.

We use the RBF kernel function for both gender and smile classifiers and the ChaLearn LAP 2016 database with binary gender and smile labels for training.

[1] Courtesy of Eyedea Recognition Ltd., http://www.eyedea.cz.

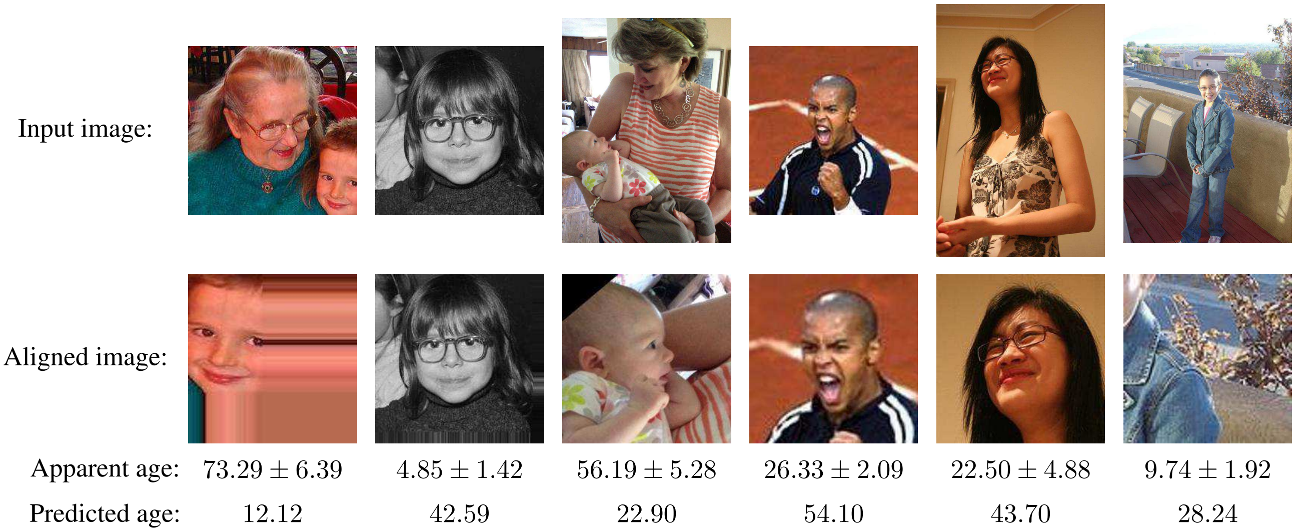

Input image:
Aligned image:
Apparent age: 73.29 ± 6.39 | 4.85 ± 1.42 | 56.19 ± 5.28 | 26.33 ± 2.09 | 22.50 ± 4.88 | 9.74 ± 1.92
Predicted age: 12.12 | 42.59 | 22.90 | 54.10 | 43.70 | 28.24

Figure 4. Examples from the validation set where our proposed method obtained the highest absolute errors.

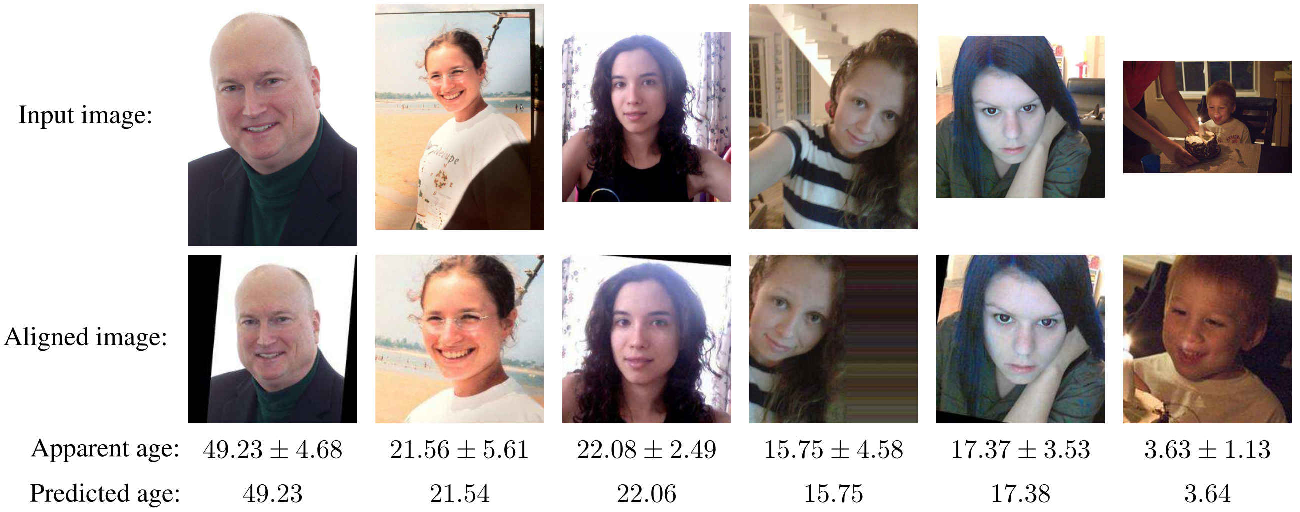

Input image:
Aligned image:
Apparent age: 49.23 ± 4.68 | 21.56 ± 5.61 | 22.08 ± 2.49 | 15.75 ± 4.58 | 17.37 ± 3.53 | 3.63 ± 1.13
Predicted age: 49.23 | 21.54 | 22.06 | 15.75 | 17.38 | 3.64

Figure 5. Examples from the validation set where our proposed method obtained the smallest absolute error.

3.3. Looking At People 2016 Challenge

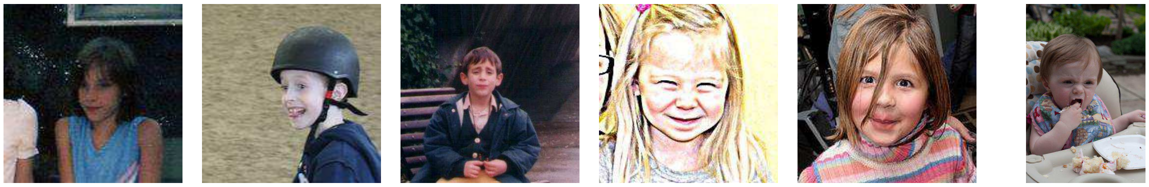
The ChaLearn LAP 2016 challenge had three tracks. The track for apparent age estimation, as well as the track for gender and smile prediction, consisted of two phases: development and test. In the next we present the results for the apparent age estimation. The final test results for gender and smile prediction will be revealed during the ChaLearn LAP 2016 workshop.

3.3.1 Development phase

In the development phase of the apparent age estimation track, only the training set annotations were released and the methods were evaluated online by submitting the predictions on the validation images to a server. The final leader board of the development phase is shown in Table 1 and reflects the performance of the last submitted predictions during the development phase for each participating team. The performance of our submission is emphasized by the bold font.

3.3.2 Test phase

In the test phase of the apparent age estimation track, the validation annotations and also the testing images were released. The testing annotations were kept secret and the teams were invited to submit their results on the testing images. The organizers announced the final ranking after the test phase. The results are summarized in Table 2. The re-



Apparent age:	9.79 ± 2.66	6.64 ± 1.64	8.50 ± 1.72	4.59 ± 1.55	5.51 ± 1.39	1.21 ± 0.40
DEX:	33.37	28.20	29.26	24.93	25.43	15.82
proposed:	8.01	5.79	7.00	6.96	4.99	1.00

Figure 6. Examples of validation images where our method significantly outperforms DEX [21].

Table 1. Ranking of all participants in the validation phase of LAP 2016 challenge on apparent age estimation. Our entry is with **bold**.

Rank	Team	ϵ-error
1	csmath	0.215327
2	xperzy	0.236426
3	**uricamic**	**0.240429**
4	OLE	0.265452
5	xbaro	0.272212
6	palm_seu	0.339589
7	really	0.345112
8	frkngrpnr	0.384979
9	rcmalli	0.417944
10	stmater	0.427319

Table 2. Ranking of all participants in the final test phase of LAP 2016 challenge on apparent age estimation. Our entry is with **bold**.

Rank	Team	ϵ-error
1	OrangeLabs	0.2411
2	palm_seu	0.3214
3	**CMP+ETH**	**0.3361**
4	WYU_CVL	0.3405
5	ITU_SiMiT	0.3668
6	Bogazici	0.3740
7	MIPAL_SNU	0.4569
8	DeepAge	0.4573

sults of the proposed method are emphasized by the bold font. Note that the results on the test phase of the methods are generally inferior to those on the development phase. There could be a different distribution of the apparent age in the test set than in the provided train and validation sets.

3.4. Discussion

3.4.1 Apparent age estimation

In the development phase, the proposed method (using a single SO-SVM classifier) got a relative improvement of 11.60% in the ϵ-error and 14.13% in MAE compared to DEX method (with a single CNN predictor, no ensemble). After the release of the validation annotations, we trained the proposed classifier in 10 fold cross-validation (on joint training and validation dataset) and got the ϵ-error reduced to 0.209 and MAE to 2.5, that is a relative improvement of 23.16% in ϵ-error and 22.38% in MAE compared to DEX.

Applying the softmax on the SO-SVM brings the improvement of 0.0042 in ϵ-error and 0.0376 in MAE on average (*i.e*. taking into account all instances in the ensemble).

Figure 4 shows the examples from the validation set where the proposed method performs the worst. In other words, we show the examples with the highest absolute errors. Note that part of these results are due to face detector failure, due to the selection of another face from the image than the one the apparent age annotation was for, due to face occlusion (glasses) or poor image quality.

In Figure 5, we show the examples from the validation set, where the proposed method achieved the smallest absolute error in apparent age prediction.

When compared with DEX (one CNN) we note that especially on the 0–10 years interval, on average, our proposed method gives much better results. In Figure 6 we show a couple of examples where DEX has errors above 14 years while our method achieves significantly lower errors, below 2 years.

3.4.2 Gender and smile prediction

We compare the proposed gender classifier to the deep learning approach by Rothe *et al*. [21], which was trained on the IMDB-WIKI dataset and outputs the softmax probability of male/female class. On the validation set of ChaLearn LAP 2016 [9] our proposed method achieves a 10.85% classification error, while the pure deep network gets 15.29% and the nearest neighbor using the same deep features reaches 16.74% classification error. We conclude that by using a non-linear SVM classifier and deep features we get a large relative improvement of 29.09% over the results achieved directly by the deep CNN predictor.

For smile prediction the nearest neighbor classifier has a 36.36% error on the validation set, while our SVM smile classifier achieves 20.97%, *i.e*. a relative improvement of

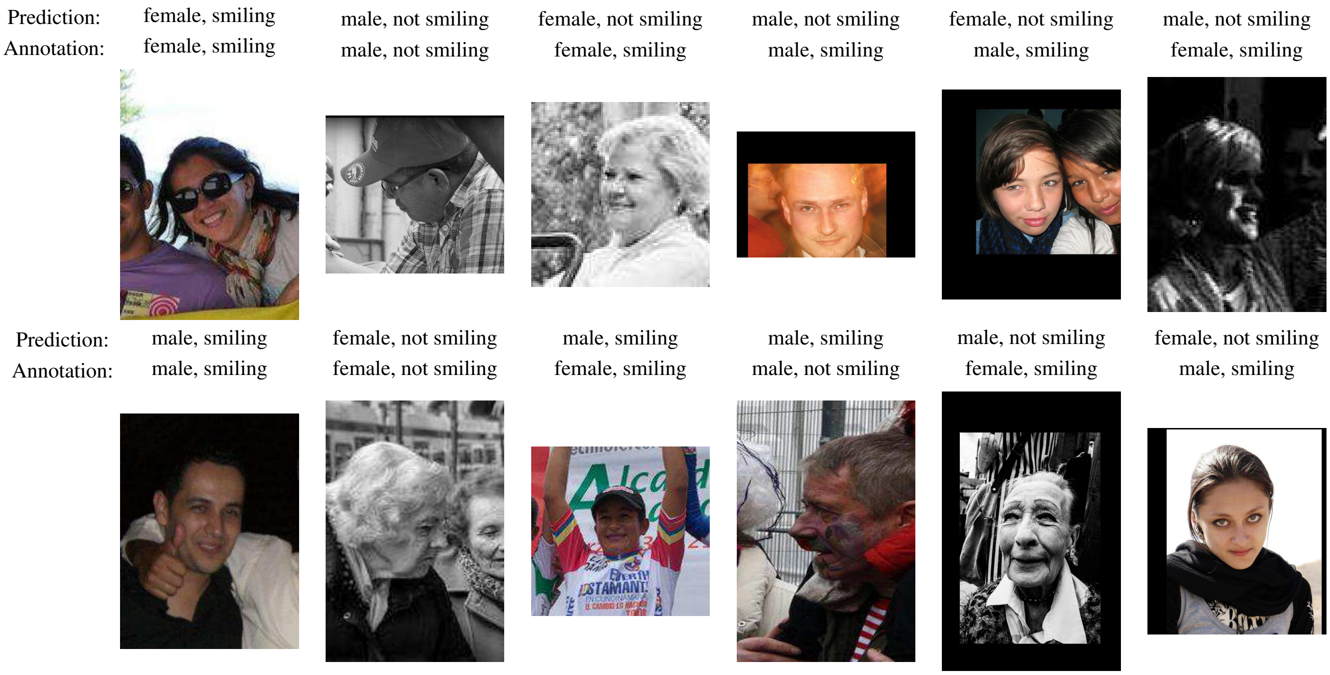


Figure 7. Examples of validation images with gender and smile predictions. First two columns show good predictions. Third and fourth column show examples where either the gender or the smile prediction is wrong. Last two columns show failure cases were no prediction matches the ground truth. Note that the bottom rightmost example has probably wrong gender annotation.

42.33%.

We got the following results for the joint evaluation of smile and gender prediction on validation data: 2.85% of examples were completely misclassified (*i.e*. both smile and gender prediction were wrong), 27.47% of examples were classified correctly in one class (*i.e*. either smile or gender prediction was correct) and 69.67% of examples were classified correctly completely (*i.e*. both smile and gender prediction were matching the annotation).

Figure 7 shows several examples with their annotations and our smile and gender predictions on the ChaLearn LAP 2016 [9] validation dataset.

4. Conclusions

In this paper we proposed the structured output SVM prediction of apparent age, gender, and smile from deep features. For apparent age prediction, our method uses an ensemble of multi-class SO-SVM predictors, which are learned from the fully annotated examples. Each multi-class SO-SVM predictor uses a softmax expected value refinement. Our experiments on apparent age, gender and smile prediction showed that our proposed approach leads to significantly better performance than the pure deep learning approach. We conclude that the best is to combine the representation power of the deep features with the robustness power of SO-SVM for prediction.

Acknowledgments

The 1st author was supported by The Czech Science Foundation Project GACR P103/12/G084 P103/12/G084 and by CTU student grant SGS15/201/OHK3/3T/13. The 4th author was supported by The Czech Science Foundation Project GACR P103/12/G084 P103/12/G084 and by CTU student grant SGS15/155/OHK3/2T/13. The 2nd, 3rd, and 5th authors were supported by The ETH General Fund (OK) and by NVIDIA with a Tesla K40 GPU donation.

References

[1] K. Antoniuk, V. Franc, and V. Hlaváč. V-shaped interval insensitive loss for ordinal classification. *Machine Learning*, 2016. 1

[2] R. Azarmehr, R. Laganiere, W.-S. Lee, C. Xu, and D. Laroche. Real-time embedded age and gender classification in unconstrained video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–65, 2015. 2

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999. 4

[4] B.-C. Chen, C.-S. Chen, and W. H. Hsu. *Computer Vision ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, chapter Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval, pages 768–783. Springer International Publishing, Cham, 2014. 1

[5] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012. 2

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference in Machine Learning (ICML)*, 2014. 2

[7] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *Information Forensics and Security, IEEE Transactions on*, 9(12):2170–2179, 2014. 1

[8] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. González, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015. 1, 2, 3

[9] S. Escalera, M. Torres, B. Martìnez, X. Baró, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar. Chalearn looking at people and faces of theworld: Face analysis workshop and challenge 2016. In *ChaLearn Looking at People and Faces of the World, CVPR workshops, CVPR*, 2016. 1, 4, 5, 7, 8

[10] V. Franc and S. Sonnenburg. Optimized cutting plane algorithm for large-scale risk minimization. *The Journal of Machine Learning Research*, 10:2157–2192, 2009. 5

[11] B.-B. Gao, X.-S. Wei, J. Wu, and W. Lin. Deep spatial pyramid: The devil is once again in the details. *arXiv preprint arXiv:1504.05277*, 2015. 2

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

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(9):1904–1916, 2015. 2

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 5

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[16] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3819–3827, 2015. 2

[17] L. Liu, C. Shen, and A. van den Hengel. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4749–4757, 2015. 2

[18] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015. 2

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2

[20] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 2, 3, 5

[21] R. Rothe, R. Timofte, and L. Van Gool. DEX: Deep EXpectation of Apparent Age From a Single Image. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015. 1, 2, 3, 4, 7

[22] R. Rothe, R. Timofte, and L. Van Gool. Some like it hot - visual guidance for preference prediction. In *CVPR*, 2016. 1

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 3

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[27] C. H. Teo, S. V. N. Vishwanathan, A. J. Smola, and Q. V. Le. Bundle Methods for Regularized Risk Minimization. *Journal of Machine Learning Research*, 11:311–365, 2010. 3, 4

[28] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005. 2, 3

[29] M. Uřičář, V. Franc, D. Thomas, A. Sugimoto, and V. Hlaváč. Multi-view facial landmark detector learned by the structured output SVM. *Image and Vision Computing*, pages –, 2016. 5

[30] J. Šochman and J. Matas. WaldBoost - Learning for Time Constrained Sequential Detection. In *The 18th IEEE Conference on Computer Vision and Pattern Recognition, CVPR'05*, pages 150–156, San Diego, CA, USA, June 2005. 5

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[32] Y. Zhu, Y. Li, G. Mu, and G. Guo. A study on apparent age estimation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015. 2