CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY IN PRAGUE

BACHELOR THESIS

ISSN 1213-2365

Face Descriptor Learned by Convolutional Neural Networks

Ondřej Holešovský

ondrej.holesovsky@fel.cvut.cz

CTU–CMP–2015–01

May 20, 2015

Available at
http://cmp.felk.cvut.cz/ftp/articles/franc/holesovsky-bc-thesis-2015.pdf

Thesis Advisor: Vojtěch Franc

Research Reports of CMP, Czech Technical University in Prague, No. 1, 2015

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Face Descriptor Learned by Convolutional Neural Networks

Ondřej Holešovský

May 20, 2015

České vysoké učení technické v Praze
Fakulta elektrotechnická

katedra řídicí techniky

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: **Ondřej Holešovský**

Studijní program: Kybernetika a robotika
Obor: Systémy a řízení

Název tématu: **Deskriptor tváří učený pomocí konvolučních neuronových sítí**

Pokyny pro vypracování:

1. Naučte hlubokou konvoluční neuronovou síť pro extrakci příznaků vhodných k rozpoznávání tváří. 2. Otestujte naučené příznaky na úloze: a) odhadu věku, b) odhad pohlaví a c) verifikace identit z obrázku tváře. 3. Porovnejte výsledky s nejlepším existujícím řešením pro danou úlohu.4. Pro učení sítě použijte knihovnu MatConvNet (url://www.vlfeat.org/matconvnet/).Testovací databáze tváří dodá vedoucí.

Seznam odborné literatury:

[1] Taigman et al.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR 2014.[2] Huang et al. : Labeled Faces in the Wild: Updates and New Reporting Procedures. Technical Report UMCS-2014-003. University of Massachutsetts. 2014.

Vedoucí: Ing. Vojtěch Franc, Ph.D.

Platnost zadání: do konce letního semestru 2015/2016

L.S.

prof. Ing. Michael Šebek, DrSc.
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 17. 12. 2014

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, dne 21.5.2015 .... Ondřej Holešovský

I would like to express my deepest gratitude to my thesis advisor, Vojtěch Franc, for priceless advice and perfect feedback about the progress of this work. Many thanks also belong to Daniel Večerka, who has been maintaining the critical computational infrastructure in the Centre of Machine Perception. Last but not least, I would like to thank my parents for their support during my studies.

Abstract

We present face descriptors created by nine and sixteen-layer deep convolutional neural networks. Performance is evaluated on identity verification and age and gender estimation tasks. Based on the neural network training scenario, a proper classifier has to be chosen in the case of identity verification. For the supervised indentity verification setting, a normalized dot product of face representation vectors outperforms the more intuitive Euclidean distance as a measure of face similarity. However, in the case of the unsupervised setting, the Euclidean distance is superior. The performance on the Labeled Faces in the Wild data set can be further improved by mirroring training images as well as by a well tuned combination of PCA and LDA projections of the face representation. Pretraining the simpler nine-layer network on the identity recognition task improves the final results in age and gender estimation. With the help of more sophisticated age and gender prediction models, both our neural networks reduce the age estimation error of the current state of the art by up to 28% on the MORPH II data set. Finally, computational performance of both neural networks is evaluated.

Prezentujeme reprezentace obrázků tváří vytvořené devíti a šestnáctivrstvou hlubokou konvoluční neuronovou sítí. Úspěšnost je hodnocena na úlohách verifikace identity a odhadu věku a pohlaví. Vhodný klasifikátor příznaků pro verifikaci identit je třeba zvolit v závislosti na postupu trénování neuronové sítě. V případě učení verifikace identity s učitelem překoná normalizovaný skalární součin intuitivnější Euklidovskou vzdálenost v přesnosti verifikace, jakožto meřítko podobnosti tváří. Avšak v případě učení bez učitele vítězí Euklidovská vzdálenost nad skalárním součinem v přesnosti verifikace. Úspěšnost na "Labeled Faces in the Wild" databázi lze dále zlepšit zrcadlením trénovacích obrázků nebo také projekcí reprezentace tváře sladěnou kombinací PCA a LDA. Předtrénování jednodušší devítivrstvé neuronové sítě na úloze rozpoznávání identit zlepší konečné výsledky odhadu věku a pohlaví. S pomocí sofistikovanějších modelů pro odhad věku a pohlaví překonají obě naše neuronové sítě nejlepší dostupné řešení na světě pro odhad věku na databázi MORPH II a to až o 28%. Nakonec hodnotíme výpočetní náročnost obou neuronových sítí.

Contents

Abbreviations

1D	one dimension(al)
2D, 3D, ...	two dimension(al), three dimension(al), two dimension(al), three dimension(al), two dimension(al), three dimension(al), two dimension(al), three dimension(al), ...
AI	artificial intelligence
CNN	convolutional neural network
RGB	red green blue, aka colour
FC	fully connected
X2	Euclidean distance
LBPPYR	pyramid of the locally binary patterns
SVM	support vector machine
SVOR-IMC	support vector ordinal regression with implicit constraints
PW-MORD	piece-wise multi-class classifier for ordinal regression
MAE	mean absolute error
PCA	principal component analysis
LDA	linear discriminant analysis
ROC	receiver operating characteristic
AUC	area under the ROC curve
CPU	central processing unit
GPU	graphics processing unit
RAM	random-access memory

1 Introduction

1.1 Thesis Assignment

Train a deep convolutional neural network suitable for feature extraction from facial images. Evaluate the extracted features on the task of: i) age estimation, ii) gender estimation and iii) identity verification based on facial images. Compare the obtained results with the current state-of-the-art.

1.2 State of the Art

The Labeled Faces in the Wild [9] (LFW) "Unrestricted with labeled outside data" protocol is currently dominated by deep convolutional neural networks. Notably the DeepFace [18] ensemble of neural networks reaches 97.35% face verification accuracy. They have trained their algorithms on a private dataset of 4 million images and 4000 identitites. Moreover, face images are being passed through a piecewise affine transformation based on 3D face modeling for alignment.

The current best algorithm on the MORPH II [13] age estimation benchmark data set is "Human vs. Machine" [8] by Hu Han, Charles Otto, and Anil K. Jain. It achieves 4.2 mean absolute age deviation, 72.5% predictions have their absolute error of the age estimate within 5 years.

Unfortunately, there is no established benchmark for gender prediction.

1.3 Thesis Structure

The second chapter of this thesis describes the convolutional neural network architectures and provides some information about training them. The features extracted by the neural networks are then classified by a prediction model which is different for each estimation/verification task. These prediction models are described in the third chapter. Data sets of images utilized for training, validation and testing our algorithms are introduced in the fourth chapter. Finally, the fifth chapter includes experiments done and their results.

Although our identity verification solutions are good, they are not yet comparable to the state of the art on LFW.

Both our trained neural networks can well classify into 16 age and gender classes (8 age classes for each gender). The vgg-16 network achieves the best result for this simple approach on the MORPH data set. With 4.09 mean absolute age deviation (MAE), it is already comparable with the current state of the art. With the help of more sophisticated age and gender prediction models (PW-MORD [1] for age estimation and a two-class linear SVM for gender prediciton), both our neural networks reduce the MAE of the current state of the art on the MORPH data set. Our Deep9 has MAE 3.90 years (improvement by 7%) and vgg-16 achieves MAE 3.04 years (improvement by 28%).

2 Main Network Architectures

This chapter aims to introduce the more complex convolutional neural network (CNN) configurations called Deep9 and vgg-16. Both of them are used for multiple classification tasks. Their performance is described in the following chapters.

2.1 Basic Facts and MatConvNet

Firstly, let us state some basic facts applying to all network models presented in this thesis. Maxpool layers are always using fields 2x2 with stride 2. Convolutional and fully connected layers are two kinds of weight layers. All networks are implemented with the help of MatConvNet toolbox [22]. Expressions used for model description:

conv-6x4-1-32
A convolutional layer with (a bank of) filters of size 6x4, input dimension 1 and output dimension 32.

FC-2048 (input 6x3-128)
A fully connected layer, output dimension 2048. Its input consists of 128 "images" of size 6x3 (a 6x3x128 matrix).

dropout 0.2
A dropout layer with 0.2 (20 %) dropout rate. It sets 20 % of passing data to zero while training.

maxpool
A max pooling layer.

ReLu
Rectified Linear Unit.

softmax
Combined softmax operator and logarithmic loss.

All these types of layers (conv/FC, ReLu, maxpool, dropout, softmax) are available as functional blocks within MatConvNet. A network can be built just by combining these blocks together.

2.2 Vgg-16 Network Architecture

This architecture was pretrained on ImageNet dataset. 'Very Deep Convolutional Networks for Large-Scale Image Recognition', Karen Simonyan and Andrew Zisserman, arXiv technical report, 2014, imagenet-vgg-verydeep-16. [16] The model is available at `http://www.vlfeat.org/matconvnet/pretrained/`.

We simply changed the last fully connected layer (different number of outputs) and retrained the whole network on our datasets. Please note that all our vgg-16 configurations were unintentionally retrained without dropout layers.

Network Configuration
Vgg-16
16 weight layers
224x224 RGB input image
conv-3x3-3-64
conv-3x3-64-64
maxpool
conv-3x3-64-128
conv-3x3-128-128
maxpool
conv-3x3-128-256
conv-3x3-256-256
conv-3x3-256-256
maxpool
conv-3x3-256-512
conv-3x3-512-512
conv-3x3-512-512
maxpool
conv-3x3-512-512
conv-3x3-512-512
conv-3x3-512-512
maxpool
FC-4096 (input 7x7-512)
dropout 0.5
FC-4096 (input 1x1-4096)
dropout 0.5
FC-464 or FC-16 etc., depends on the specific task
soft-max

Table 2.1 Configuration of vgg-16 network. There is a ReLu placed after every conv layer and also after both FC-4096 layers. Please note that all our vgg-16 configurations were unintentionally retrained without dropout layers.

2.3 Deep9 Network Architecture

Deep9 network is inspired by vgg-16 configuration a bit. Its first 5 convolutional layers use 3x3 filters. However, randomly initialized network does not converge. In order to solve this problem, we initialize the first 5 conv layers from the pretrained 3x3 vgg-16 filters by means of ”copy and paste” method, leaving the remaining weight layers initialized randomly.

Network Configuration
Deep9
9 weight layers
60x40 grayscale input image
conv-3x3-1-32
conv-3x3-32-32
maxpool
conv-3x3-32-64
conv-3x3-64-64
maxpool
conv-3x3-64-128
conv-5x3-128-128
FC-2048 (input 6x3-128)
dropout 0.5
FC-2048 (input 1x1-2048)
dropout 0.5
FC-464 or FC-16 etc., depends on the specific task
soft-max

Table 2.2 Configuration of Deep9 network. There is a ReLu placed after every conv layer and also after both FC-2048 layers.

3 Prediction Models

Prediction models are algorithms processing CNN features in a way to solve a particular classification or estimation task. Unless stated otherwise, CNN features are the input of the last fully connected (FC) layer.

3.1 Age and Gender Estimation

3.1.1 Two-Class Linear SVM for Gender Estimation

We predict the gender information from CNN features by a simple two-class linear SVM classifier [14]. The classifier

$$y = h(\mathbf{x}, \mathbf{w}, b) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{3.1}$$

returns $y = -1$ for a female and $y = 1$ for a male face. The trainable parameters comprise of $\mathbf{w}, b$. $\mathbf{x}$ is a feature vector.

3.1.2 Piece-Wise MORD for Age Estimation

In order to obtain age estimations from CNN feature vectors, we employ Piece-wise Multi-class Classifier for Ordinal Regression [1]. Firstly, we split the ages $y \in Y$ into age groups Y_z. These age groups are separated by Z cut labels $\hat{y}_1 < \hat{y}_2 < \cdots < \hat{y}_Z$. We can define PW-MORD by plugging the linear combination coefficient

$$\alpha(y, z) = \frac{y - \hat{y}_z}{\hat{y}_{z+1} - \hat{y}_z} \tag{3.2}$$

into the classifier formula

$$h(\mathbf{x}, \mathbf{W}, \mathbf{b}) = \operatorname*{arg\,max}_{z \in Z} \operatorname*{arg\,max}_{y \in Y_z} \Big(\langle \mathbf{x}, \mathbf{w_z}(1 - \alpha(y, z)) + \mathbf{w_{z+1}} \alpha(y, z) \rangle + b_y \Big), \tag{3.3}$$

where $\mathbf{W}, \mathbf{b}$ are the parameters to be trained and $\mathbf{x}$ is a feature vector.

3.2 Identity Verification

Identity verification methods guess if two feature vectors ($\mathbf{x_1}$ and $\mathbf{x_2}$) denote the same person or not.

3.2.1 Euclidean Distance

This is probably the most intuitive metric. We may assume that two close features result in an identity match, whereas two distant features reveal different identities. The distance d,

$$d = \sqrt{\sum_{i=1}^{D} (x_{1i} - x_{2i})^2}, \tag{3.4}$$

where D is the number of dimensions, $\mathbf{x_1}$ and $\mathbf{x_2}$ are the feature vectors being compared. Having the distance computed, the classification algorithm (same/different) consists simply in comparing the distance with a previously learned constant threshold. Moreover, a ROC curve (e.g. false versus true positive rate) of the verification algorithm can be drawn by varying this threshold.

3.2.2 L2 Normalization, Dot Product

A dot product can help us distinguish if 2 feature vectors point to a similar or different direction.

$$d(\mathbf{x_1}, \mathbf{x_2}) = \langle \mathbf{x_1}, \mathbf{x_2} \rangle \tag{3.5}$$

We may use d in the same way as the Euclidean distance.

Furthermore, the feature vectors can be L2 normalized before computing their dot product.

$$d(\mathbf{x_1}, \mathbf{x_2}) = \frac{\langle \mathbf{x_1}, \mathbf{x_2} \rangle}{\|\mathbf{x_1}\|^2 \|\mathbf{x_2}\|^2} \tag{3.6}$$

This approach is suggested in [18] (DeepFace).

3.2.3 PCA + LDA

The feature vectors can be further preprocessed before using them for classification as outlined above. We can at first reduce their dimension by Principal Component Analysis (PCA). The compressed features can then be projected into a more discriminative subspace by means of Linear Discriminant Analysis (LDA). [5] The LDA projection matrix is trained on a set of identity labeled, PCA compressed features. See the Experiments chapter for more information.

4 Datasets

The sections "Age and Gender Estimation Data Sets", "Data Preprocessing" and "Data Augmentation" have been mostly written by the thesis supervisor Vojtěch Franc.

4.1 Age and Gender Estimation Data Sets

We utilize two databases for the age and gender estimation task.

The MORPH II database [13] is a collection of US police mugshots. The images have a fixed resolution 200x240px and each of them contains a single subject. The images have been taken under controlled lighting conditions. Most of the faces are frontal and with subjects showing a neutral facial expression. The images are annotated with the true age and gender of the depicted subjects. The MORPH database has become the standard benchmark used for evaluation of age estimation systems.

The LPIP database [1] is composed of three standard face databases and additional images downloaded from the Internet. The three databases are the Labeled Faces in the Wild [9], PAL [12] and PubFig [10]. Because the sample of young and old age categories is not sufficiently representative in none of the three databases they have been augmented by images downloaded from the Internet to make the age distribution closer to uniform. The faces exhibit large variation in races, facial expressions, background clutter, resolution and the quality of images in the sense of focus or motion blur. The faces are near frontal with the yaw and pitch angles between -30 and 30 degrees. The images are labeled by age and gender estimated by a set of human annotators.

To sum up, the MORPH database contains near perfect input images with ground truth annotation. In contrast, the LPIP database represents challenging "in-the-wild" images with inconsistent human-created annotation. Table 4.1 shows some basic statistics. The age distribution is depicted in Figure 4.1. Sample images drawn from both datasets are in Figure 4.2 and Figure 4.3.

	male	female	total	annotation	description
MORPH II database	46,645	8,489	55,134	ground truth	Police database of mugshots. Controlled conditions, resolution 200x240.
LPIP database	25,295	23,534	48,829	human	Images in the wild. Uncontrolled, mixed quality. LFW + PubFig + Internet download + PAL.

Table 4.1 Databases used in benchmark.

4.1.1 Split to Training/Validation/Testing Part

The images were split randomly three times into training, validation and testing part in the ratio 60/20/20. We made sure that images of the same identity never appear

[1] Courtesy of Eyedea Recognition s.r.o www.eyedea.cz

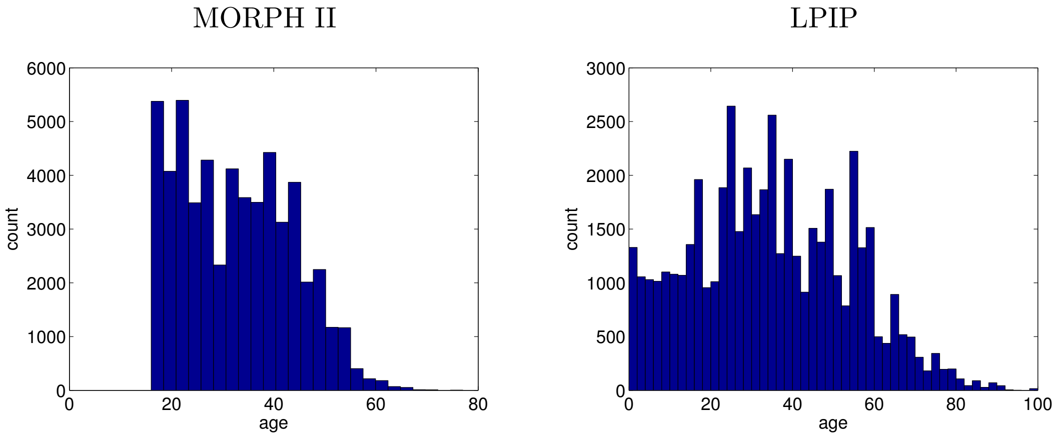

Figure 4.1 Age distribution in the MORPH II and the LPIP database, respectively.

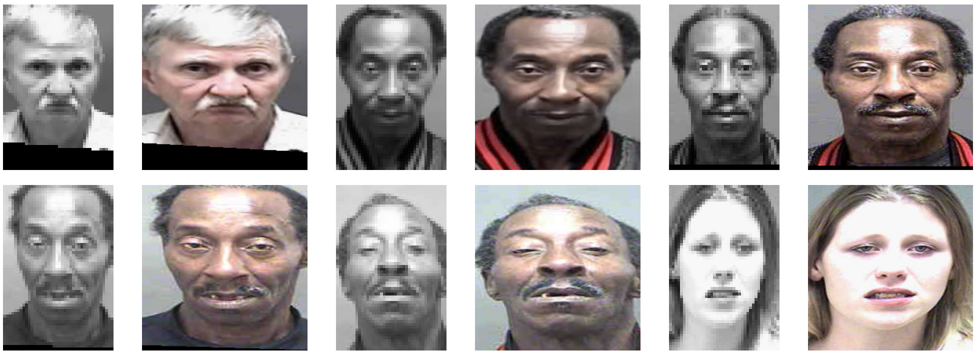

Figure 4.2 Sample images from MORPH dataset. Both 60x40 gray and 224x224 RGB versions side by side.

in the different parts simultaneously. The training and validation parts are used for training the parameters and model selection. The test results are mean and standard deviations computed over the three test splits.

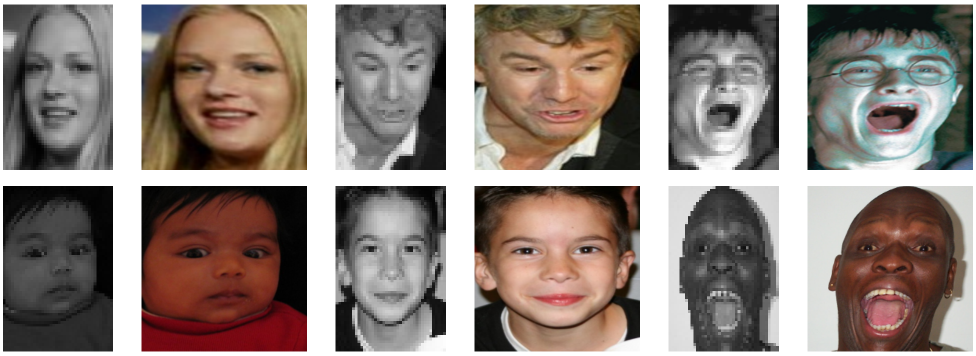

Figure 4.3 Sample images from LPIP dataset. Both 60x40 gray and 224x224 RGB versions side by side.

4.2 Identity Verification

4.2.1 LFW

Labeled Faces in the Wild (LFW) [9] data set is de facto the standard benchmark for face verification algorithms. The data set contains 13233 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. Overall, there are 5749 different people in the data set.

The LFW test protocol suggests reporting performance as 10-fold cross validation using splits they have once randomly generated. Each test split contains 300 matched pairs and 300 mismatched pairs of images.

We test the performance of our algorithms on LFW mostly under the protocol called "Unrestricted with labeled outside data". This means that labeled outside data can be used for training, as long as there is no overlap with the LFW dataset. Our baseline vgg-16 network pretrained just on ImageNet is tested under the "Unsupervised" protocol. Any training data are permitted in the "Unsupervised" protocol, as long as they are not annotated with identity or same/different information. Further information is available in the New Technical Report (http://vis-www.cs.umass.edu/lfw/lfw_update.pdf).

4.2.2 FaceScrub + PubFig

Our dataset for identity recognition and verification training is built by merging FaceScrub [7] (http://vintage.winklerbros.net/facescrub.html) and PubFig [10] datasets. All identities contained also in LFW are removed. FaceScrub dataset has URLs published for 107,818 images of 530 people. However, many of these images are no more available at their specified URLs. The original PubFig dataset has 58,797 images of 200 people, also in the form of listed URLs. Furthermore, some overlaps in identities appear during the merge of FaceScrub with PubFig and removal of LFW. Some identities get more samples but many have to be deleted. The resulting dataset consists of about 82,000 images and 464 identities. Figure 4.4 shows some sample images.

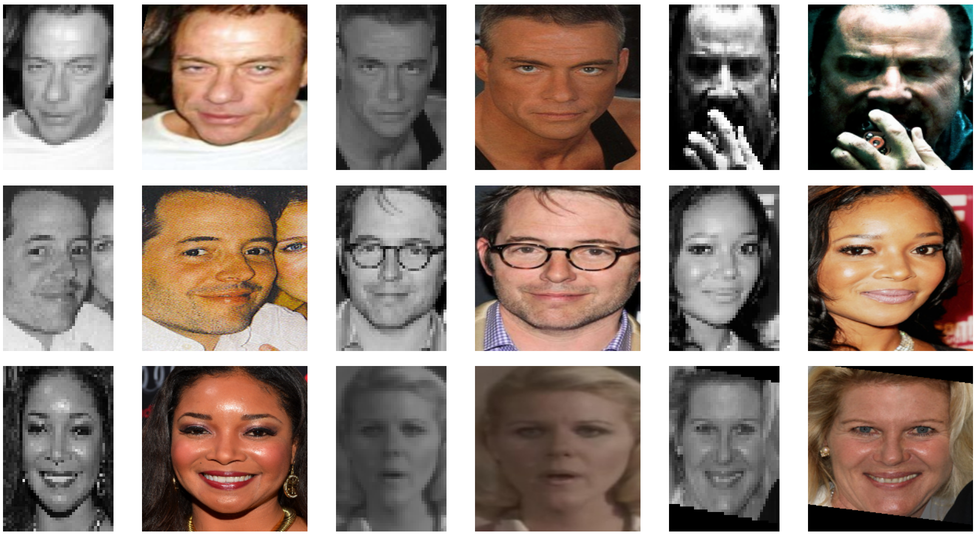

Figure 4.4 Sample images from FaceScrub + PubFig dataset. Both 60x40 gray and 224x224 RGB versions side by side.

4.3 Data Preprocessing

The feature representation of the facial images were computed as follows. We first localized the faces by a commercial face detector [2] and consequently applied a Deformable Part Model based detector [20] to find facial landmarks like the corners of eyes, mouth or tip of the nose. The found landmarks were used to transform the face by an affine transform into its canonical pose. Finally, the features were extracted from the canonical face image. We considered two resolutions of the canonical face: 60x40 and 224x224 pixels. We used two feature descriptions. First, the pyramid of the locally binary patterns (LBPPYR) proposed in [17]. The LBPPYR descriptor applied on 60x40 image is a sparse $n = 159,488$-dimensional binary vector serving as an input of linear classifier (either two-class SVM or the PW-MORD classifier). Second, the features learned by the deep convolutional neural network.

4.4 Data Augmentation

The deep networks require a large set of training examples in order not to overfit. We enlarged the training set by generating synthetic facial examples.

We generated a set of facial bounding boxes for each input face in the database. The bounding boxes were derived from the base bounding box found by the detector. In particular, the based bonding box was perturbed by applying in-plane rotations (+5 and -5 degrees) and scaling (1.05 and 0.95). Consequently, we extracted the canonical face from each generated bounding box by the process described above. Finally, we doubled the number of the canonical faces by mirroring them around the vertical axis.

[2] Courtesy of Eyedea Recognition s.r.o www.eyedea.cz

5 Experiments

This chapter presents the experiments done and their results.

5.1 Age and Gender Estimation

In order to train a CNN for the simultaneous age and gender estimation task, we decide to discretize the ages into 8 groups. As a result, we convert the task to a 16 class classification problem (8 age groups times 2 genders). That means the output dimension of the last fully connected network layer is set to 16. It is followed by a softmax loss layer.

The abbreviations of the evaluation metrics and linear classification models presented in the tables:

MAE The mean absolute deviation between the age estimated by computer and the age in the database. The database age is either the ground truth (MORPH) or human estimate (LPIP).

CS5 The portion of images with the absolute error of the age estimate not higher than 5 years.

maleAsFemale The portion of images where male is erroneously estimated as female.

femaleAsMale The portion of images where female is erroneously estimated as male.

total The combined classification error for the 16 class classification task.

SVOR-IMC Support vector ordinal regression with implicit constraints [4].

PW-MORD Piece-wise linear Multi-class classifier for Ordinal Regression [1].

SVM Two-class linear Support Vector Machine classifier [21].

5.1.1 Training Data Augmentation and Pretraining Effects

We have completed a few experiments to discover which training procedure is the best for each network. The following tables present the validation errors of the age and gender estimation experiments. The validation errors are calculated on the LPIP validation set after every epoch during each CNN training on LPIP. We take the 16 class discretized prediction and compare it with the precise labels. The standard deviation is calculated only for the better results when we have trained the networks on 3 splits.

We can infer from table 5.1 that Deep9 network gives better performance when pretrained firstly on the identity recognition task and then trained on LPIP than if it is trained directly on LPIP from (mostly) random initialization. Vgg-16, on the contrary, achieves better performance when not pretrained on the ID task.

Deep9 benefits from augmented training data (see table 5.2). Vgg-16 results are nearly the same for both the augmented and non-augmented training set.

Configuration	Net Validation Errors			
	total	maleAsFemale	femaleAsMale	MAE
Deep9	49.53	11.61	10.41	6.9
Deep9 ID pretrained	45.92 ± 2.11	9.47 ± 1.23	10.51 ± 0.46	6.47 ± 0.12
Vgg-16	37.83 ± 1.59	6.53 ± 0.77	7.03 ± 1.05	5.30 ± 0.36
Vgg-16 ID pretrained	42.91	7.76	6.88	6.0

Table 5.1 Effects of previous pretraining on the identity recognition task. LPIP data set.

Configuration	Net Validation Errors			
	total	maleAsFemale	femaleAsMale	MAE
Deep9	49.67	12.04	10.34	7.0
Deep9 augmented	45.92 ± 2.11	9.47 ± 1.23	10.51 ± 0.46	6.47 ± 0.12
Vgg-16	39.20	6.90	6.06	5.3
Vgg-16 augmented	37.83 ± 1.59	6.53 ± 0.77	7.03 ± 1.05	5.30 ± 0.36

Table 5.2 Effects of the training data augmentation. LPIP data set.

5.1.2 The Results and the State of the Art Comparison

Test errors are calculated on the test set. The features are extracted from images by a CNN and fed into a prediction model. Precise age predictions from Piece-Wise MORD, for example, are compared to the precise labels. Gender is determined by the two-class linear SVM on the test set.

Training is performed on the augmented data sets. Deep9 network is pretrained on the identity recognition task, vgg-16 only on ImageNet.

	MAE	CS5 [%]	note
[8] Han Hu. vs. M.	4.2	72.4	Morph II (20569 subj, 78207 img)
[3] Chang H. Rank.	6.1	56.3	Morph (55608 images)
[6] Guo KPLS	4.2	NA	Morph (55608 images)
SVORIMC	5.00(±0.03)	64.2 (±0.11)	60x40; LBPPYR features
PwMord	4.67 (±0.04)	68.1 (±0.53)	60x40; LBPPYR features
Deep9-lpip	11.6 (±0.06)	NA	60x40; Deep9 trained on LPIP
PwMord-Deep9-lpip	5.38 (±0.01)	61.2 (±0.3)	60x40; Deep9 trained on LPIP
Deep9	4.80 (±0.10)	NA	60x40; Deep9 trained on MORPH
PwMord-Deep9	3.90 (±0.04)	75.2 (±0.6)	60x40; Deep9 trained on MORPH
vgg-16-lpip	6.60 (±0.10)	NA	224x224; vgg-16 trained on LPIP
PwMord-vgg-16-lpip	4.58 (±0.17)	68.2 (±1.8)	224x224; vgg-16 trained on LPIP
vgg-16	4.09 (±0.09)	NA	224x224; vgg-16 MORPH trained
PwMord-vgg-16	3.04 (±0.08)	84.4 (±0.9)	224x224; vgg-16 MORPH trained

Table 5.3 MORPH II database: Age estimation.

We conclude that the more complex vgg-16 network outperforms the simpler Deep9 network. Except complexity, the higher dimensional (224x224) colour images the network processes can be another reason for the better vgg-16 results. They carry more information than the 60x40 grayscale images being fed into Deep9 network. What is more surprising, our experiments show that data augmentation has only negligible impact on the age and gender classification errors. The choice of a right prediction model greatly affects estimation accuracy of both the age (see tables 5.3, 5.5) and gender

	maleAsFem.	fem.AsMale	note
SVM	2.99 (±0.21)	3.33 (±0.28)	60x40; LBPPYR features
Deep9-lpip	20.7 (±0.88)	18.5 (±1.85)	60x40; Deep9 trained on LPIP
SVM-Deep9-lpip	6.45 (±0.41)	5.86 (±0.69)	60x40; Deep9 trained on LPIP
Deep9	0.93 (±0.12)	5.44 (±0.30)	60x40; Deep9 trained on MORPH
SVM-Deep9	2.24 (±0.12)	2.63 (±0.42)	60x40; Deep9 trained on MORPH
vgg-16-lpip	6.99 (±0.37)	24.2 (±0.27)	224x224; vgg-16 trained on LPIP
SVM-vgg-16-lpip	4.50 (±0.70)	4.30 (±0.33)	224x224; vgg-16 trained on LPIP
vgg-16	0.54 (±0.10)	2.29 (±0.67)	224x224; vgg-16 trained on MORPH
SVM-vgg-16	1.05 (±0.38)	1.07 (±0.26)	224x224; vgg-16 trained on MORPH

Table 5.4 MORPH II database: Gender estimation.

	MAE	CS5 [%]	note
SVORIMC	9.18(±0.19)	39.8 (±0.80)	60x40, LBPPYR features
PwMord	7.27 (±0.13)	56.7 (±0.67)	60x40; LBPPYR features
Deep9	6.47 (±0.12)	NA	60x40; Deep9 trained on LPIP
PwMord-Deep9	5.76 (±0.17)	65.1 (±0.8)	60x40; Deep9 trained on LPIP
vgg-16	5.30 (±0.36)	NA	224x224; vgg-16 trained on LPIP
PwMord-vgg-16	4.30 (±0.11)	74.3 (±0.5)	224x224; vgg-16 trained on LPIP

Table 5.5 LPIP: Age estimation.

(tables 5.4, 5.6). PwMord and SVM outperform the native 16 class classifier for each CNN configuration. The SVM makes the gender estimation errors (maleAsFemale and femaleAsMale) more balanced. What's more, the CNN features provide a better input for PwMord and SVM than the LBPPYR features. This can be also seen in all the four tables.

Both our CNNs improve on the state of the art of age estimation on MORPH, namely from 4.2 MAE [8] to 3.04 and 3.9 for vgg-16 and deep9 respectively (table 5.3). We also obtain better results on LPIP in age estimation than PwMord and LBPPYR with MAE 7.27, precisely 5.76 for deep9 and 4.30 for vgg-16 (table 5.5). The SVM gender classification is also significantly improved by Deep9 and vgg-16 features (tables 5.4, 5.6).

	maleAsFemale	femaleAsMale	note
SVM	11.72 (±1.37)	11.38 (±2.57)	60x40; LBPPYR features
Deep9	9.47 (±1.23)	10.51 (±0.46)	60x40; Deep9 trained on LPIP
SVM-Deep9	8.28 (±0.78)	7.63 (±1.38)	60x40; Deep9 trained on LPIP
vgg-16	6.53 (±0.77)	7.03 (±1.05)	224x224; vgg-16 trained on LPIP
SVM-vgg-16	4.50 (±0.82)	4.07 (±1.43)	224x224; vgg-16 trained on LPIP

Table 5.6 LPIP: Gender estimation.

5.2 Identity Classification and Verification

Performance of our algorithms in this chapter is evaluated under two LFW protocols. ”Unsupervised” and the supervised ”Unrestricted with labeled outside data”. Most of this chapter is devoted to the supervised setting, only the last subsection of this chapter presents the ”Unsupervised” results.

5.2.1 CNN Trained on an Identity Classification Task

In the supervised setting, we train both Deep9 and vgg-16 CNN architectures by letting them classify into 464 classes (the amount of identities in FaceScrub + PubFig dataset). When a network has been well trained (a low FaceScrub + PubFig validation error), we extract the features of all LFW images by the network and employ one of the prediction models. However, PCA and LDA training is done only on FaceScrub + PubFig dataset. As a result, we have used from LFW only the testing samples for benchmarking our algorithms.

5.2.2 Effects of Training Data Changes

We have done some experiments to understand how various changes of the training data set impact the classification/verification performance. All these experiments are realized on vgg-16 CNN. Figure 5.1 shows that augmented training data significantly improve both validation and LFW test errors. However, 2x data augmentation (original and mirrored image) seems to cause the largest improvement so further augmentation probably would not help a lot.

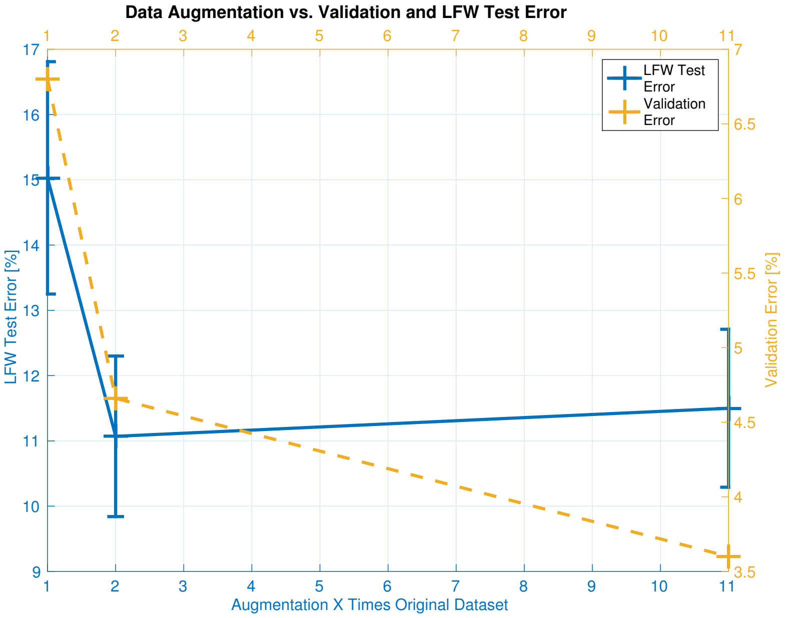

Figure 5.1 Identity classification (validation) and LFW test errors are affected by training data augmentation.

The number of identities represented in the data set seems to affect performance (Figure 5.2). The classification task becomes harder with more identities in the data set so in accordance with our expectations the validation error rises. Although the LFW performance improves with more identities, adding much more of them may be of no use.

According to DeepFace [18], the amount of image samples per identity has a significant impact on the performance. We have tried to verify this finding on our own (Figure 5.3). It seems that altering the amount of samples per identity has no effect

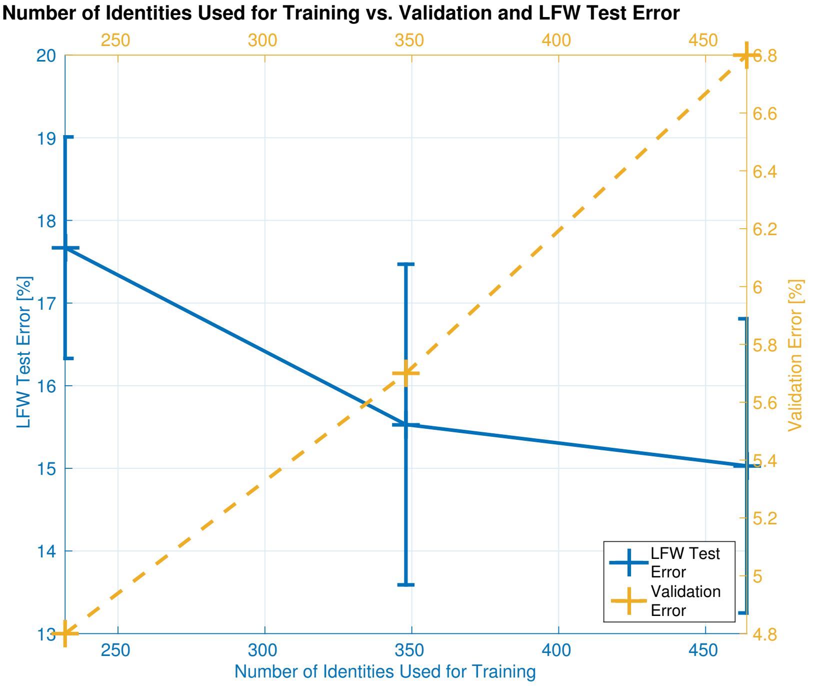

Figure 5.2 Identity classification (validation) and LFW test errors are affected by the number of identities included in the training data.

on LFW performance. However, we have too few samples per identity (about 150 on average, DeepFace has about 1000) so that the experiment is not much valuable.

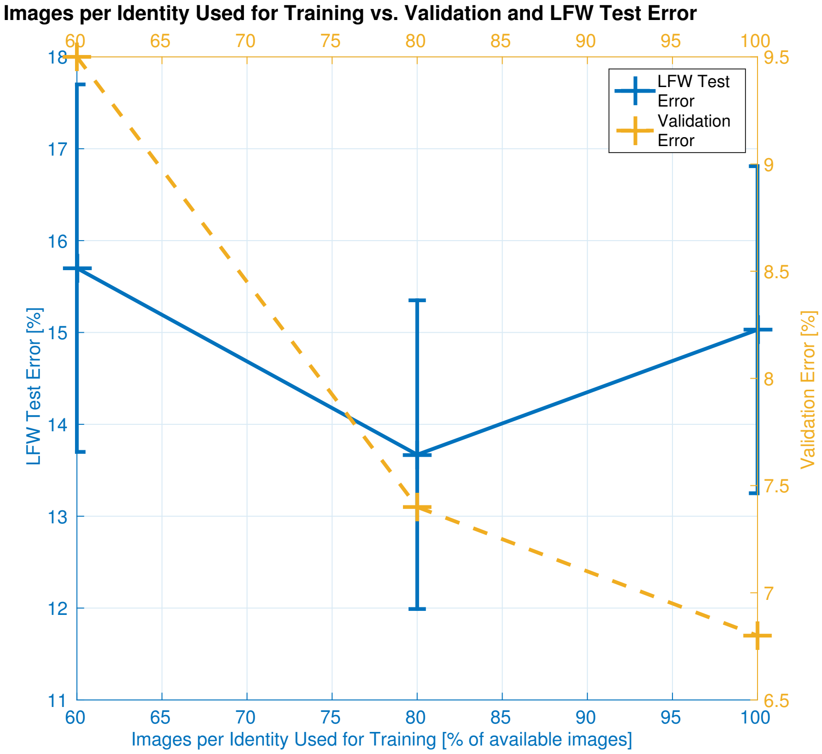

Figure 5.3 Identity classification (validation) and LFW test errors are affected by the amount of image samples per each identity in the training dataset. Please note that 100 % per ID is about 150 images in our case which may be too few for such evaluation. DeepFace [18], for example, utilizes up to 1000 images per identity in their training dataset.

5.2.3 PCA+LDA Tuning

PCA and LDA projections are trained on 4096 dimensional vgg-16 extracted features of non-augmented (original) images. The output dimensions of PCA and LDA projections have to be set in a way which minimizes the validation error.

Figures 5.4 and 5.5 show how the FaceScrub + PubFig validation error is affected by various combinations of PCA and LDA output dimensions. The Euclidean distance version achieves the optimum at PCA dimension 81 and LDA dimension 70. Further raising the dimensions brings no improvement. In contrast, the dot product version reaches its optimum at PCA dim. 391 and LDA dim. 280. The error may decrease further but we cannot substantially increase the dimensions as we have only 400 identities in the training data set. LDA output dimension cannot be larger than the number of training identities.

We can say that better results are achieved when the LDA output dimension is closer to the PCA output dimension.

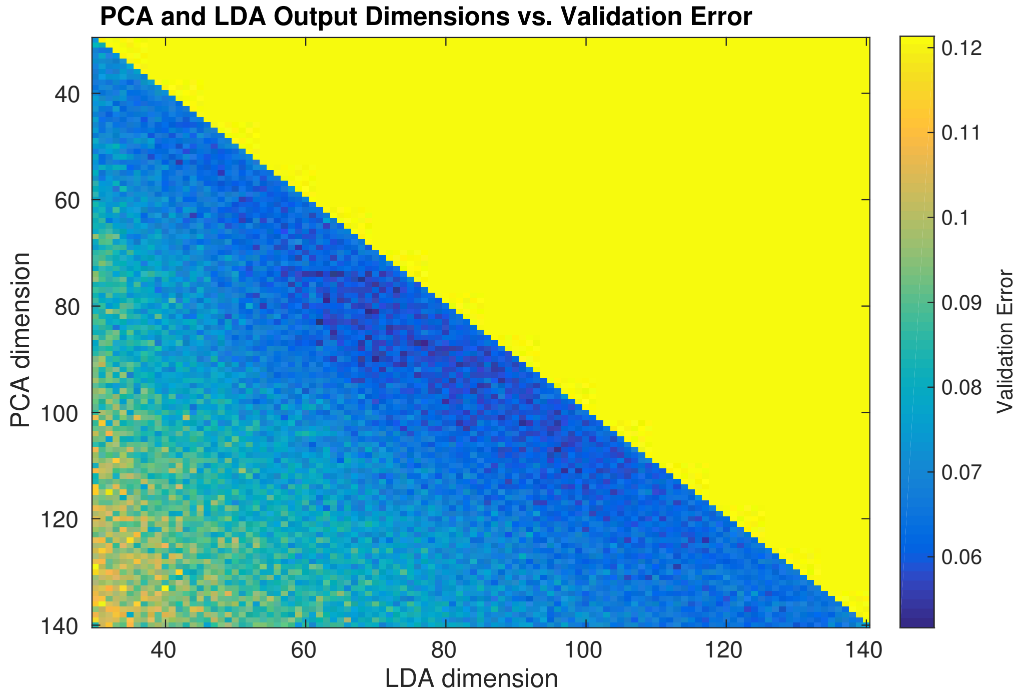

Figure 5.4 PCA + LDA + Euclidean distance identity verification.

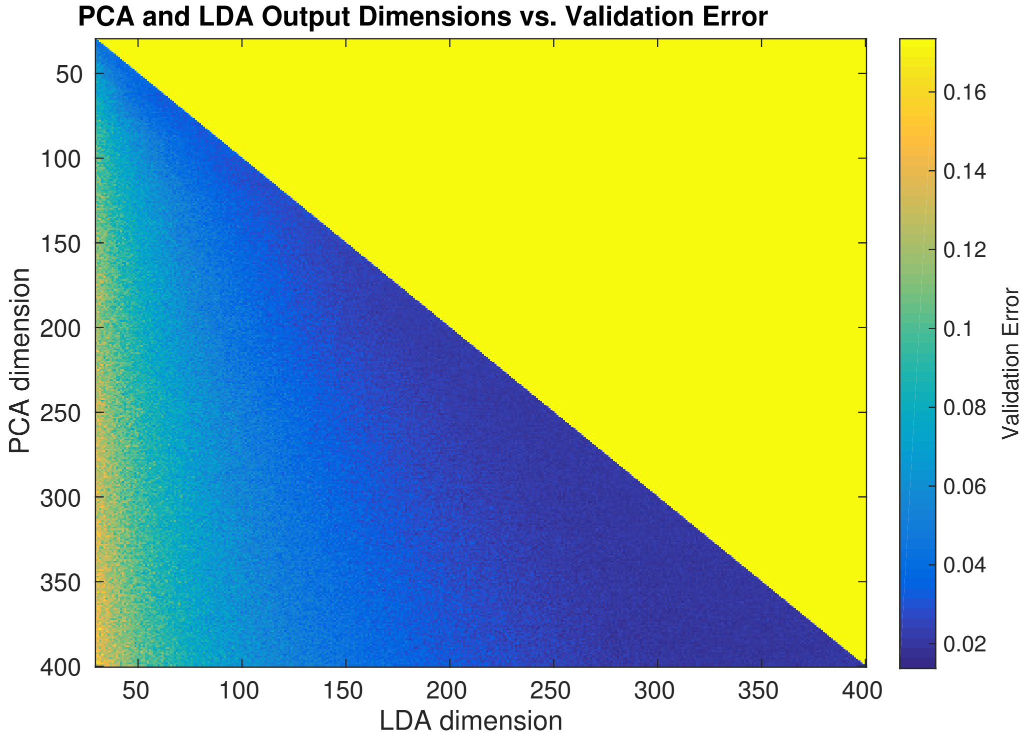

Figure 5.5 PCA + LDA + L2 normalization + dot product identity verification.

5.2.4 The Supervised Results and the State of the Art Comparison

The following tables and ROC curves compare the results of our algorithms with each other (table 5.8, figure 5.6) and also our best algorithm with the state of the art on LFW (table 5.7, figure 5.7).

Algorithm	Success Rate
Vgg-16, PCA 391, LDA 280, L2 norm, dot	0.9060 ± 0.0126
DeepFace-ensemble [18]	0.9735 ± 0.0025
Simile classifiers [11]	0.8414 ± 0.0041
Eigenfaces, original [19]	0.6002 ± 0.0079

Table 5.7 LFW ”Unrestricted with labeled outside data” test results, state-of-the-art comparison. Number written after ”PCA” or ”LDA” is the output dimension of the projection. ”dot” denotes the dot product, ”L2 norm” is L2 normalization.

Configuration: Error Type	Error
Deep9: FaceScrub + PubFig validation	7.7%
Deep9, X2: LFW	$27.63 \pm 1.97\%$
Deep9, L2 norm, dot: LFW	$\mathbf{17.07 \pm 1.31}\%$
Vgg-16: FaceScrub + PubFig validation	3.6%
Vgg-16, X2: LFW test	$17.13 \pm 2.10\%$
Vgg-16, PCA 81, LDA 70, X2: LFW	$15.63 \pm 2.20\%$
Vgg-16, L2 norm, dot: LFW	$11.50 \pm 1.21\%$
Vgg-16, PCA 391, LDA 280, L2 norm, dot: LFW	$\mathbf{9.40 \pm 1.26}\%$

Table 5.8 Deep9 and vgg-16 results. FaceScrub + Pubfig Validation and LFW test errors on various prediction models are presented. Number written after ”PCA” or ”LDA” is the output dimension of the projection. ”X2” denotes Euclidean distance, ”dot” the dot product, ”L2 norm” is L2 normalization. All LFW results are in compliance with the ”Unrestricted with labeled outside data” protocol.

We again conclude that the more complex vgg-16 network outperforms the simpler Deep9 network. However, even our best solution (vgg-16) cannnot compete with one of the best peer-reviewed algorithms to date (DeepFace [18]) under the ”Unrestricted with labeled outside data” setting.

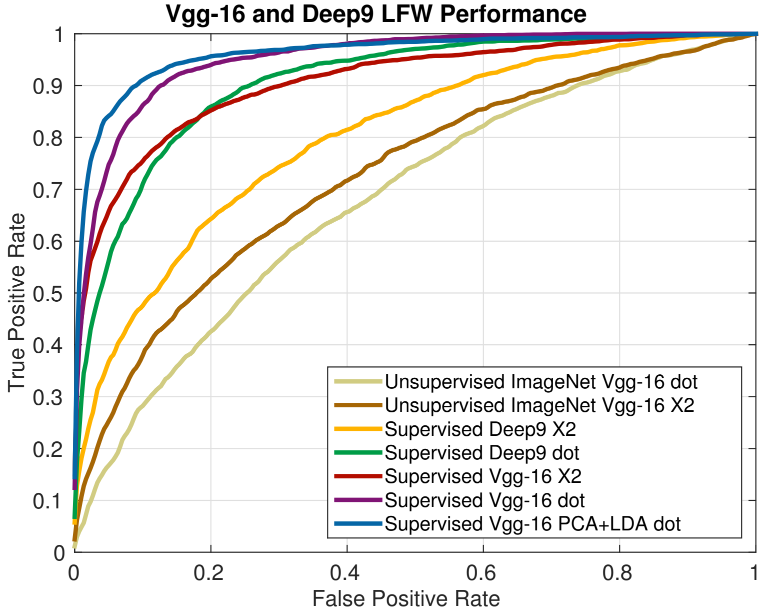

Figure 5.6 ROC curves on LFW dataset of both networks.

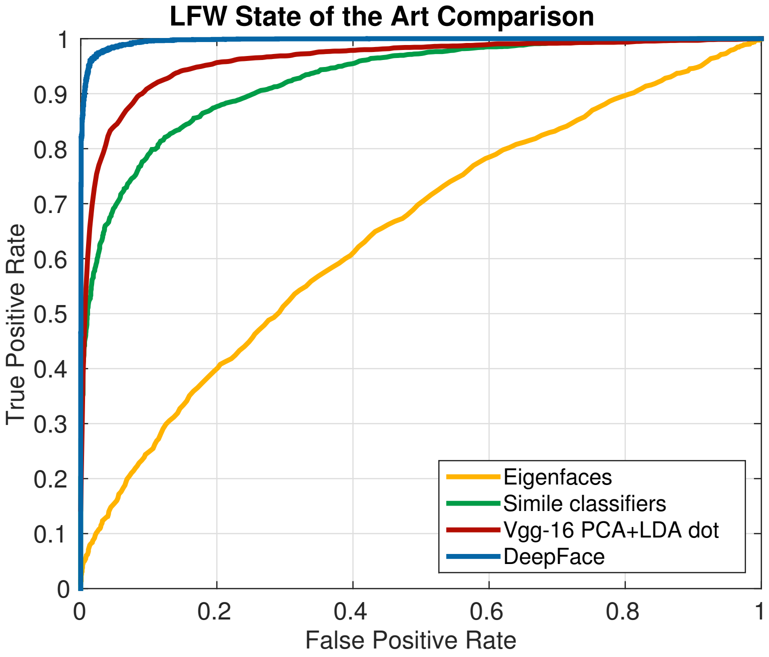

Figure 5.7 Comparison of the ROC curves of our best solution so far (Vgg-16 PCA+LDA dot) with other algorithms on LFW. LFW protocol: "Unrestricted with labeled outside data".

5.2.5 The LFW Unsupervised Protocol Results

We have also done some tests under the ”Unsupervised” LFW protocol. Networks trained on ImageNet and on the age and gender estimation task comply with the ”Unsupervised” setting. Vgg-16 configuration trained on the age and gender estimation task and equipped with the Euclidean distance (X2) prediction model is quite comparable to the second best ”Unsupervised” LFW algorithm to date called MRF-MLBP [2]. However, the comparison is probably not completely fair. Many LFW images are included in our age and gender training data set LPIP. See figure 5.8 for the ROC curves and table 5.9 for the AUC results.

Algorithm	AUC
ImageNet Vgg-16 dot	0.6803
ImageNet Vgg-16 X2	0.7270
Age Gender Vgg-16 dot	0.8296
Age Gender Vgg-16 X2	0.8980
LHS [15]	0.8107
MRF-MLBP [2]	0.8994
Pose Adaptive Filter (PAF) [23]	0.9405

Table 5.9 LFW ”Unsupervised” test results, state-of-the-art comparison. The table includes our results (vgg-16) and the three best state-of-the-art solutions. AUC is the area under the ROC curve.

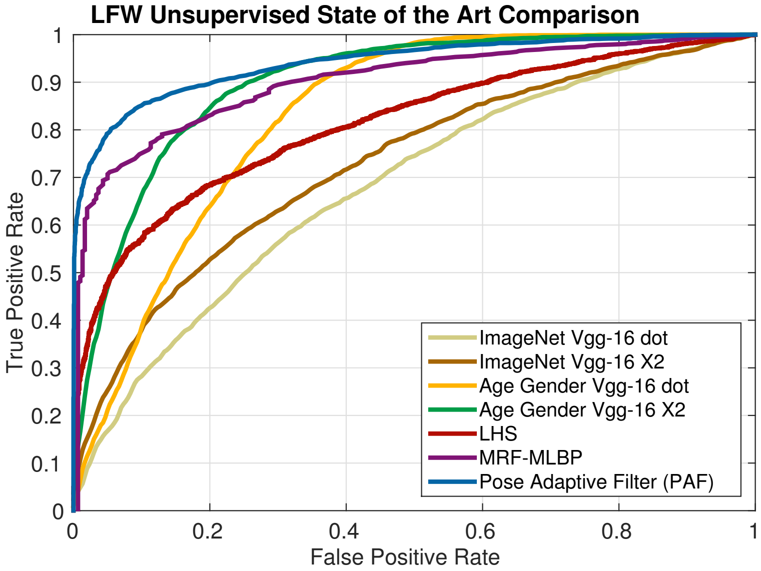

Figure 5.8 Comparison of the ROC curves of ”Unsupervised” algorithms on LFW. Our results (vgg-16) and the three best state-of-the-art solutions are shown.

5.3 CNN Computational Performance

Classification accuracy of an algorithm is not the only important information to know when one decides how to solve a real-life problem. In most cases, the computational power demands of an algorithm are also important. Table 5.10 shows how fast our two

CNN models are on a computer with Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz (32 cores), 126 GB RAM, GPU NVIDIA GeForce GTX TITAN Black (6 GB memory, 3.5 compute capability). In all the performance tests, the input images are either directly preloaded in RAM or are at least already cached from a hard drive.

Deep9 configuration is not only much faster, it also requires less GPU memory. Deep9 fits even in 2 GB, whereas vgg-16 requires 6 GB of memory space on the GPU when being trained. It is better to process more images at once (use larger batches), if possible. The speed then rises considerably on the GPU. Implemeting the network models in C/C++ directly, for example, may further boost performance.

The CPU performance can be seen in the table 5.11. The Deep9 performance on the CPU is even better for smaller batches than on the GPU. In contrast, the more complex vgg-16 processes only 2 images per second on the CPU regardless of the batch size.

Algorithm	Training Speed	Feature Extraction Speed
Vgg-16	10 (batch size 10)	64 (bs 20), 31 (bs 2), 20 (bs 1)
Deep9	280 (bs 100), 130 (bs 10)	3800 (bs 100), 970 (bs 20), 98 (bs 2), 50 (bs 1)

Table 5.10 Computational performance (speed) on the GPU of both Deep9 and vgg-16 network models in processed images per second. Batch size (bs) also affects the speed.

Algorithm	Feature Extraction Speed
Vgg-16	2 images/s (bs 20, 10, 2, 1)
Deep9	140 (bs 100), 140 (bs 20), 138 (bs 10), 110 (bs 2), 86 (bs 1)

Table 5.11 Computational performance (speed) on the CPU of both Deep9 and vgg-16 network models in processed images per second. ”bs” is batch size.

6 Conclusions

Two deep convolutional neural networks have been tested.

- **Deep9:** 9 weight layers, 60x40 grayscale input image, smaller, faster, less accurate results.
- **Vgg-16:** 16 weight layers, 224x224 RGB input image, larger, slower, more accurate results. Developed and pretrained on ImageNet by Karen Simonyan and Andrew Zisserman [16].

Good results have been achieved on the Labeled Faces in the Wild data set in two protocols. We obtain 0.8980 area under the ROC curve in the ”Unsupervised” and 90.60% accuracy in the ”Unrestricted with labeled outside data” protocols. It should be noted, however, that none of our algorithms has seen the LFW training data with identity or same/different labels. Only in the case of our best ”Unsupervised” solution, many LFW images have been included in the training data set (LPIP). The following same/different identity prediction models have been tested.

- **Euclidean distance:** Works best in the ”Unsupervised” setting, when networks are trained on ImageNet or the age and gender estimation task.
- **L2 normalization and dot product:** Serves best in the supervised setting (”Unrestricted with labeled outside data”).

Results of both prediction models in the supervised setting are improved by the tuned PCA and LDA face descriptor mappings. Augmentation of training images affects the classification error. However, simple mirroring is sufficient. Further artificial training set enlargement brings only a mild improvement.

Both neural networks can well classify into 16 age and gender classes (8 age classes for each gender). The vgg-16 network achieves the best result for this simple approach on the MORPH data set. With 4.09 mean absolute age deviation (MAE), it is already comparable with the current state of the art. With the help of more sophisticated age and gender prediction models (PW-MORD [1] for age estimation and a two-class linear SVM for gender prediciton), both our neural networks reduce the MAE of the current state of the art on the MORPH data set, which is [8] (Han, Human vs. Machine) with MAE 4.2 years. Our Deep9 has MAE 3.90 years (improvement by 7%) and vgg-16 achieves MAE 3.04 years (improvement by 28%).

Pretraining the simpler Deep9 network on the identity recognition task improves the final results in the age and gender estimation.

All experiments have been carried out in Matlab and MatConvNet toolbox [22]. Computations have mostly run on a GPU.

Bibliography

[1] Kostiantyn Antoniuk, Vojtech Franc, and Vaclav Hlavac. Mord: Multi-class classifier for ordinal regression. In *ECML/PKDD (3)*, pages 96–111, 2013. 3, 7, 13, 24

[2] Shervin Rahimzadeh Arashloo and J Kittler. Efficient processing of mrfs for unconstrained-pose face recognition. *Proc. Biometrics: Theory, Applications and Systems*, 2013. 22

[3] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplane ranker with cost sensitivities for age estimation. In *CVPR*, 2011. 14

[4] Wei Chu and S. Sathiya Keerthi. New approaches to support vector ordinal regression. In *In Proceedings of the 22nd international conference on Machine Learning (ICML)*, pages 145–152, 2005. 13

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, NY, USA, 2nd edition, 2001. 8

[6] Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer Vision and Pattern Recognition*, 2011. 14

[7] S. Winkler. H.-W. Ng. A data-driven approach to cleaning large face datasets. In *Proc. IEEE International Conference on Image Processing (ICIP), Paris, France*, pages 343–347, Oct 2014. 11

[8] Hu Han, Charles Otto, and Anil K. Jain. Age estimation from face images: Human vs. machine performance. In *International Conference on Biometrics (ICB)*, 2013. 3, 14, 15, 24

[9] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 3, 9, 11

[10] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009. 9, 11

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009. 20

[12] M. Minear and D. Park. A life span database of adult facial stimuli. *Behavior Research Methods, Instruments & Computers*, 36(4):630–633, 2004. 9

[13] Karl Ricanek and Tamirat Tesafaye. 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, April 2006. 3, 9

[14] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002. 7

[15] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie. Local higher-order statistics (lhs) for texture categorization and facial analysis. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VII*, ECCV'12, pages 1–12, Berlin, Heidelberg, 2012. Springer-Verlag. 22

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 24

[17] Sören Sonnenburg and Vojtěch Franc. Coffin: A computational framework for linear svms. In *Proceedings of the 27th Annual International Conference on Machine Learning (ICML 2010)*, Madison, USA, June 2010. Omnipress. 12

[18] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1701–1708. IEEE, 2014. 3, 8, 16, 18, 20

[19] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, Jun 1991. 20

[20] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of facial landmarks learned by the structured output SVM. In Gabriela Csurka and José Braz, editors, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556, Porto, Portugal, February 2012. SciTePress - Science and Technology Publications. 12

[21] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems. Wiley, New York, New York, USA, 1998. 13

[22] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564, 2014. 4, 24

[23] Dong Yi, Zhen Lei, and Stan Z. Li. Towards pose robust face recognition. In *CVPR*, pages 3539–3545. IEEE, 2013. 22