BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2020
http://excel.fit.vutbr.cz

Automated human recognition from image data

Lukáš Dobiš

Abstract
This paper describes an approach for automated human recognition by using convolutional neural networks (CNN) to perform facial analysis of persons face in image data. The predicted biometric indicators are following: age, gender, facial landmarks and facial expression. CNN architectures with pretrained weights for each task are described. Age estimation CNN has new weights trained and freezed, then has added new LSTM layers into its architecture. New LSTM layers are trained and tested on newly created video data set. Solution for human recognition inference with single image and time series variants, in form of script with interconnected CNNs is explained and its inference speed performance supports further proposed expansion plans for live video inference.

Keywords: Human classification — Computer vision — Deep learning

Supplementary Material: GitHub repository

*xdobis01@stud.feec.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Automated human recognition as a technology is used widely in human-computer interaction, the security and surveillance industry, demographic research, commercial development, mobile application, service robotics and video games [44]. Use of computer vision for human recognition, has greatly increased with introduction of deep learning methods based on CNN.

State-of-art neural networks often focus only on one task, so combining multiple of them using their strengths, leads to composition of networks each responsible for different task. This way used networks can have their architecture or weights changed at will. So when better architecture for one task is discovered it can be used as replacement without need to retrain whole composition of networks. With this approach it is also easier to train, test and evaluate networks, since independent tasks have often their own benchmark data sets.

This paper applies multiple pretrained convolutional neural networks to detect face, its landmarks and to further predict person age, gender and facial expression. These are important descriptors in automated human recognition. Since each human has unique appearance and leads different lifestyle, these descriptors precise prediction creates a challenge for current algorithms. Aim of this work is to create composition of CNNs oriented on human recognition, which will be capable of real-time inference in video, with improvement in prediction by utilizing temporal information.

Age estimation being hardest task is main focus. So in addition to pretrained weights has used age prediction network new weights trained and freezed. Then is modified with additional LSTM layers which are independently trained and tested on new data set created for this purpose. Results show that adding LSTM layers has improved age prediction performance.

Human recognition is therefore solved with distinct CNNs interconnected to create automated human face detection and subsequent age, gender and emotion classification from images. Overall lightweight single image inference time on GPU is 0.1038 second. Modified solution for video inference, needs 20 images for age prediction and has inference time 0.1181s.

By using several lightweight precise CNNs, the proposed solution is fast, accurate and highly modular for further improvements.

2. Background and related works

2.1 Object detection

Most specialized detectors like face detectors are based on general purpose object detection architectures [38],

which use feature extraction networks pretrained on large data sets like ImageNet [12], PASCAL2007 [14] or MS COCO [23].

As an object detection consist of two steps. First step is finding parameters of so called bounding box which fits over detected object and second detected object classification. To well known approaches belong: Faster R-CNN[31], Deconvolutional Single Shot Detector (DSSD)[16], You Only Look Once detector v.3 (YOLOv3) [30], RetinaNet[24] and CenterNet[41]. Approaches are described in order of their first appearance.

Faster R-CNN [31] is fourth iteration on region proposing CNN approach. Its one of two stage methods, which use two separate parts of network, one to propose regions for object location and second to classify object. This approach is the most accurate but large number of region proposals affects inference speed. Other mentioned approaches are one-stage methods so localizing and classifying is done at once by same parts of network. This greatly improves inference speed but at cost of accuracy mainly for smaller objects.

Deconvolutional SSD [16] approach draws feature maps from feature extractor at differing depth. Bounding boxes have spatial geometric parameters regressed and class probabilities estimated at each drawn depth level.

YOLOv3 [30] divides image into grid of cells, where for each cell all possible bounding boxes are predicted. All this takes place in form of fully convolutional network, so YOLO is extremely fast and invariant to the input image size.

Relatively new is RetinaNet [24] approach, its innovation lies in using different type of loss function for classification rather than classic cross entropy. Hard to classify samples of rare class are often in low number and are overwhelmed by easy samples contribution to loss, which then dominates the gradient. RetinaNet loss function called Focal loss solves this by increasing importance of correcting misclassified samples. In terms of speed and accuracy RetinaNet stands in the middle, Figure 1.

New modern approach CenterNet [41] treats objects as a single point, center point of its bounding box. It uses keypoint estimation to find object center points and regresses other object properties like size, orientation and pose. CenterNet architecture claims to be simpler, faster and more accurate then other bounding box detectors. But currently to authors knowledge CenterNet based facial detectors outperformed RetinaNet based architectures only in terms of speed.

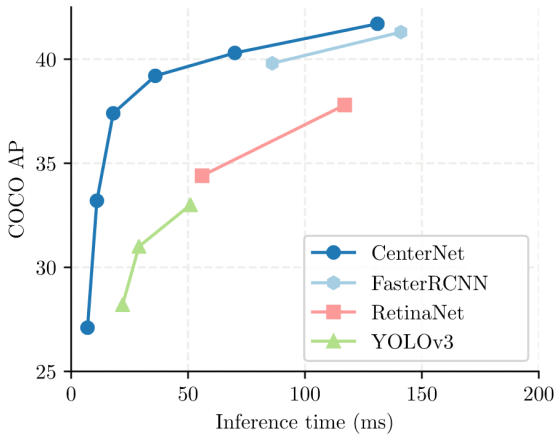


Figure 1. Speed-accuracy trade-off on COCO validation for some of the state-of-the-art real-time detectors [41].

Since this work should be capable of real-time inference, the most accurate approach Faster R-CNN was not chosen. Therefore facial detector used in this work is based on second most accurate approach RetinaNet. This facial detector is called RetinaFace[13] and is examined in subsection (3.1).

2.2 Age estimation

Current approaches treat age as regression metric or as ordinal property to be classified. Ordinal regression (OR) CNNs have been proven to outperform metric regression CNN [25].

In the field of machine learning OR has given a way to extend classification algorithms to solve regression tasks like age estimation, by reformulating problem to utilize multiple binary classification tasks. Where binary tasks represent range of values which output can be classified. First attempts used perceptrons and support vector machines [11][10]. Then came unifying general reduction framework from OR to binary classification [22]. Later this framework has been used for CNN estimating age and had proven to be at state-of-art level [25]. It portrayed OR problem with K ranks as $K-1$ binary classification problems. Where k ($k \in (1,2,..,K-1)$) task predicted whether age label of a face on image is greater than tasks rank r_k. All tasks share intermediate layers but have distinct weights in output layer. While successful at the time this approach had flaw in its classifier inconsistency. There was no assurance that tasks do not contradict each other. For example one task predicts age not above 20 and higher rank task predicts age above 30.

Modification of [25] called Ranking-CNN trains a series of CNNs and aggregates their output to predict age label of a given face image. This improves the predictive performance in comparison with a single CNN

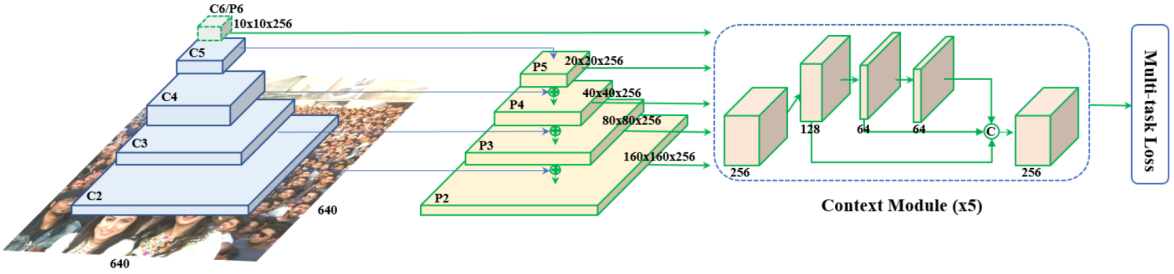


Figure 2. The RetinaFace neural network architecture [24].

with multiple binary outputs. But there is disadvantage in form of considerable increase in training complexity and still does not address classifier inconsistency [8].

Another approach utilizing binary classifiers for OR is siamese CNN architecture. With only single output neuron, comparisons are made between input image and multiple meticulously selected anchor images. From this comparisons an output rank label is computed, this is one of the solutions to class inconsistency [28].

Example of non OR approach is All-in-one CNN, training for more general face attributes analysis tasks (face detection,gender prediction, age estimation, etc.) improves the overall performance of metric regression, by sharing lower-layer parameters [29].

Different non OR approach CNN with cascades was designed to classify face images into age groups followed by metric regression modules for more accurate age estimation [7].

Last mentioned method is OR Consistent Rank Logits (CORAL) framework. This methods solves class inconsistency problem and addresses training data set uneven class frequency. Method will be thoroughly examined in subsection (3.2), as it is used in this paper [5].

2.3 Gender classification

There are various method to classify gender, this survey will mention methods using RGB images only. First methods used Support vector machine classifiers on image intensities [37]. Same concept using image intensities, used Adaboost instead [2], or applied Weber's Local texture descriptor [36]. Other methods used image intensity together with shape and texture features in combination with mutual information [27]. Another combination was Local binary pattern (LBP) and an Adaboost classifier [34]. First well established method using CNN was [20], others followed [3], [4]. Last mentioned is used by this paper, and described in subsection (3.4).

2.4 Emotion classification

Overview of emotion recognition methods can be found in [19]. Some of the often cited works are [21] and [4]. Latter is used in this paper. Temporal information has also been proved to help better classify emotions [15].

3. Used architectures

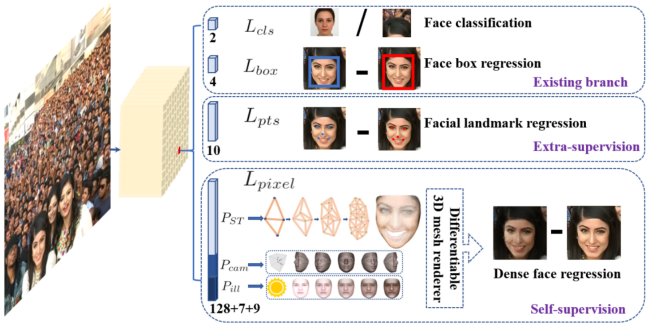
There are 4 CNNs used by this paper, first CNN is the most crucial as its output is input to all other CNNs. Its role is to detect persons face. Other 3 CNNs use this face as input to predict their own predictions.

3.1 RetinaFace network

RetinaFace is face detector based on RetinaNet architecture [13]. It broadens goals of face detection from just traditional bounding box parameters prediction to include also prediction of 5 facial landmarks. As loss function it uses multitask focal loss function (1),which takes into account face alignment, pixel-wise face parsing and 3D dense correspondence regression , Figure 3.

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) + \lambda_3 p_i^* L_{pixel} \quad (1)$$

Figure 3. RetinaFace one-stage pixel-wise face localization employs extra-supervised and self-supervised multi-task learning in parallel with the box classification and regression branches [13]



First element $L_{cls}(p_i, p_i^*)$ is face classification binary softmax loss, with p_i being probability of anchor

box having a face and p_i^* equal to 1 for positive anchor and 0 for negative anchor. Second element $L_{box}(t_i,t_i^*)$ is box regression loss equal to $\mathrm{R}(t_i - t_i^*)$ where R is $smooth - L_1$ from [31], parameters $t_i = [t_x,t_y,t_w,t_h]$, $t_i^* = [t_x^*,t_y^*,t_w^*,t_h^*]$ are predicted and ground truth bounding box geometric transformation parameters associated with positive anchor. Third element is $L_{pts}(l_i,l_i^*)$ is facial landmark regression loss with l_x and l_y coordinates parameters for 5 landmarks and each predicted l_i and ground truth l_i^* landmarks. Their transformation and loss is computed same as with box regression loss. Fourth element L_{pixel} is Dense Regression Loss and has equation (2). Dense Regression Loss is pixel-wise difference of rendered 2D face $\mathscr{R}(\mathscr{D}_{P_{ST}},P_{cam},P_{ill})$ and ground truth face. W and H are width and height of anchor crop $I_{i,j}^*$. $\mathscr{R}$ is 3D mesh renderer used with shape and texture parameters P_{ST} to project a colouredmesh $\mathscr{D}_{P_{ST}}$ onto 2D image plane with camera parameters P_{cam} and illumination parameters P_{ill}.

$$L_{pixel} = \frac{1}{W*H}\sum_{i}^{W}\sum_{j}^{H}||\mathscr{R}(\mathscr{D}_{P_{ST}},P_{cam},P_{ill})_{i,j} - I_{i,j}^*||_1 \tag{2}$$

Last loss-balancing parameters λ_1,λ_2,λ_3 are weights with values 0.25,0.1 and 0.01 this means that bounding box and landmark loss is deemed more important than dense correspondence regression loss, Figure 3.

Mathematical apparatus to predict P_{ST} uses mesh decoder with graph convolution method based on fast localised spectral filtering formulated as Chebyshev polynomial truncated at order K, with each order evaluated at scaled Laplacian $\hat{L}$. This filtering operation uses sparse matrices and is extremely fast but its thorough explanation is above extent of this paper so for full mesh decoder explanation one should read RetinaFace paper [13].

RetinaFace architecture uses feature pyramid network with its outputs flowing through context modules to shared loss head, Figure 2.

Independent context module is applied to each of the feature pyramid levels to increase the receptive field and enhance the rigid context modeling power. In lateral connections the deformable convolution network strengthens the non-rigid context modeling capacity.

With confidence threshold of 0.05 are most of the bounding box predictions filtered out. Non maximum suppression is applied with intersection over union threshold equal to 0.4 on top 400 confidence score bounding boxes. Each overlapping bounding box contribution to final prediction location is weighted by his confidence score.

3.2 CORAL model

Consistent Rank Logits model is CNN with ordinal responses. Its authors provide theoretical guarantees of class consistency, well defined generalization bounds, task-specific importance weighting, while accomplishing lower number of parameters to be trained in comparison to other state-of-art CNNs. These guarantees are for full review in original paper [5].

Age rank label y_i is extended into $K-1$ binary labels $y_i^{(1)},...,y_i^{K-1}$, where $y_i^{(k)} \in \{0,1\}$ indicates if y_i exceeds rank r_k (3). This indication is for latter purposes, defined as indicator function $\mathbb{1}\{\cdot\}$ which if inner condition is true returns 1, otherwise returns 0.

$$y_i^{(k)} = \begin{cases} 1 & y_i > r_k \\ 0 & y_i \leq r_k \end{cases} \tag{3}$$

So instead of single ordinal age label, has CORAL anotation form of binary vector where ones start at index of original age label ordinal value 4. This allows for training simple CNN with $K-1$ binary classifiers in output layer. They share weight parameter but have independent bias which effectively deals with class inconsistency and lowers number of parameters.

$$h(\mathbf{x_i}) = r_q \tag{4}$$

$$q = 1 + \sum_{k=1}^{K-1} f_k(\mathbf{x_i}) \tag{5}$$

From binary classifier tasks responses a predicted rank r_q is obtained in equation (4). And binary label vector index q is equal to one plus sum of predictions $f_k(x_i) \in \{0,1\}$ where k is index of binary classifier in output layer (5). All predictions f_k must reflect ordinal information and at same time be rank-monotonic. Rank-monotonic rule for ordinal values can be simply explained for age case as moving one ordinal value to higher index must increase its actual non ordinal metric value and same must hold for moving in opposite direction (6) [5].

$$f_1(x_i) \geq f_2(x_i) \geq ... \geq f_{K-2}(x_i) \geq f_{K-1}(x_i) \tag{6}$$

Let $\mathbf{W}$ denote weight parameters excluding the bias units of output layer. Let penultimate layer, whose output is denoted as $g(\mathbf{x}_i,\mathbf{W})$, shares a single weight with all neurons of output layer. To $g(x_i,\mathbf{W})$ are then added $K-1$ independent bias units. Together $\{g(\mathbf{x}_i,\mathbf{W}) + b_k\}_{k=1}^{K-1}$ create inputs to the corresponding binary classifiers in the output layer. Then predicted empirical probability for binary classifier task k is defined in equation (7). Where s is logistic sigmoid function (8).

$$\hat{P}(y_i^{(k)} = 1) = s(g(\mathbf{x}_i,\mathbf{W}) + b_k) \tag{7}$$

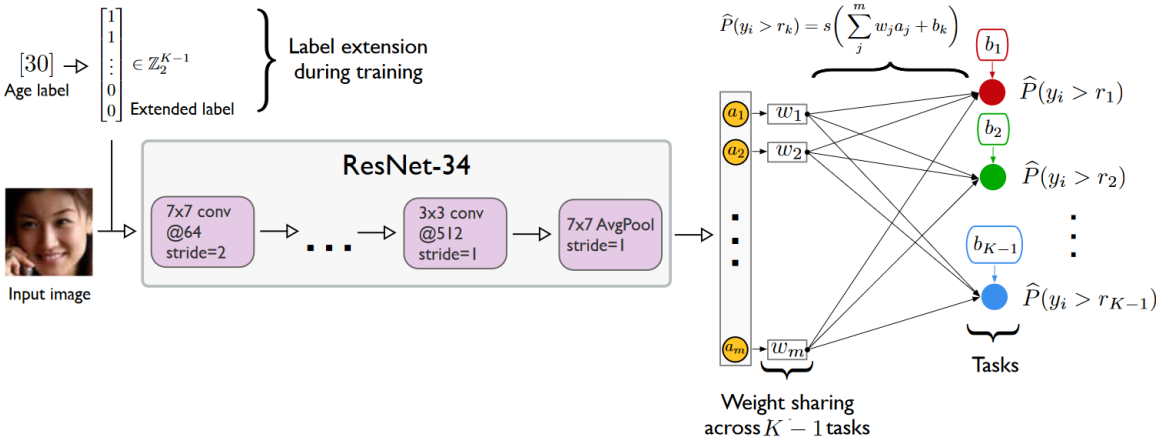


Figure 4. CORAL model architecture [5]

$$s(z) = \frac{1}{(1+exp(-z))} \tag{8}$$

In training CORAL minimized loss function is weighted CE of $K-1$ binary classifiers (9). Where $\lambda^{(k)}$ denotes the weight of loss associated with k-th classifier. For rank prediction, binary labels $f_k(\mathbf{x}_i)$ are obtained from equation (10) [5].

$$L(\mathbf{W},\mathbf{b}) = -\sum_{i=1}^{N}\sum_{k=1}^{K-1}\lambda^k[\log(s(g(\mathbf{x}_i,\mathbf{W})+b_k))y_i^{(k)} + \log(1-s(g(\mathbf{x}_i,\mathbf{W})+b_k))(1-y_i^{(k)})] \tag{9}$$

$$f_k(\mathbf{x}_i) = \mathbb{1}\{\hat{P}(y_i^{(k)}=1) > 0.5\} \tag{10}$$

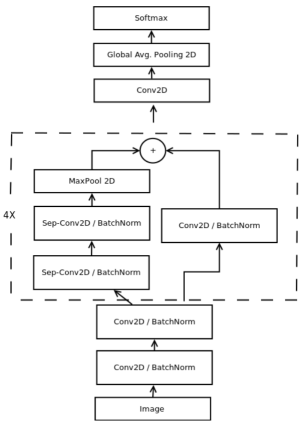
3.3 Proposed modification of CORAL model

Long-short time memory cell (LSTM) is a recurring neural network architecture used for time sequences of data [17]. It has very high rate of success and is one of main drivers of current AI research growth [45]. By adding LSTM layers to current CORAL model, network should be able to learn to extract temporal information to improve its performance. In CORAL architecture in Figure 4, are features after propagating through ResNet-34, concatenated into 2048 feature long 1D tensor. This tensor then acts as input to last linear layer which predicts results of binary classification tasks. By adding two LSTM layers to first process a time series of concatenated 1D tensors of multiple images, final LSTM output 2048 feature long tensor should be enriched with temporal information to provide provide better prediction in last linear layer.

3.4 Gender and emotion classification

Gender classification network [4] classifies into woman and man labels. For this it utilizes fully-convolutional neural network for binary classification architecture. Emotion is classified similarly, but using same architecture for multi-class classification. It can classify emotion into angry, disgust, fear, happy, sad, surprise and neutral labels [4]. Both networks take greater crop then bounding box face because networks were trained to utilize semantic information about hair to deliver more accurate prediction, mainly for gender. Architecture components are in Figure 5.

Figure 5. Architecture mini-Xception, used in gender,emotion classification networks [4]



4. Implementation

Implementation of CNNs has been written in programming language Python. All networks except one modified CORAL variant, had their weights and necessary code for network initialization downloaded from github repositories of CNNs authors. RetinaFace and CORAL networks were implemented using PyTorch framework[26], while gender and emotion networks were implemented with Keras API[9] using TensorFlow framework[1].

4.1 Script initialization

Main script is Recognition-net.py that takes as input directory of images to be analyzed and directory to save analyzed images with prediction metadata. Other optional inputs are RetinaFace base architecture

and weights (default ResNet-50 or MobileNet-0.25), CORAL weights training data set (AFAD [25],CACD[6], default MORPH[32],UTK[40], new UTK (1-100)) and processing unit (default CPU,GPU).

Script contains functions and classes from authors scripts which are used to create network model and load weights. Script starts with reading user inputs and based on user input initializes each CNN.

4.2 Face detection

Image is passed to RetinaFace network which returns bounding boxes in form of top left and bottom right coordinates and their probability scores. Third output are parametrized landmarks that are decoded into image coordinates from image size. Boxes with low probability are rejected and overlapping boxes are united using non-maximum suppression.

4.3 Facial analysis

Information about top left and bottom right corner is used to crop raw image. Cropped image of face is resized into 128x128 and centered. In this form is image passed to CORAL model which returns vector of age probabilities, this vector is thresholded by 0.5 (10) and then is summed into estimated age label. Bottom age limit of picked data set of CORAL weights is added to estimated age label. This is the final age prediction.

Next bounding box location is used again to crop out face, but this time is used area larger than detected bounding box. It takes 50% wider and 66% longer face area than original bounding box. This cropped face image is converted to gray scale. Gray and RGB image variants are then normalized and RGB variant is resized into 48x48 size and gray variant into 64x64 size. RGB face image goes into gender classification network and gray face image is input to emotion classification network. Both networks output a vector of label probabilities where prediction is label with maximum probability.

All recognized faces with their predictions are drawn onto raw image and saved to inputted save directory. Along with analyzed image the following data is stored in .txt file with same name: recognized faces bounding boxes metadata with predicted age, gender and emotion.

4.4 CORAL LSTM modification

Videos of subjects with age range 0-100 were downloaded from Youtube channel SoulPancake [46]. Each video had around around 50 or more unique individuals with their age displayed in bottom left corner. Each person had their scene extracted and first 20 images of scene were used to create one sequence per person. Training data set has 68 sequences, and testing data set has 35 sequences. Majority of individuals are not present in both data sets. In few cases where individual is present in both data sets, there is age difference of around 3-4 years.

Age estimation CORAL network had new weights trained on UTK data set[40]. Using same setting as authors with 200 epochs, Adam optimizer and learning rate of 0.0005 [5]. These weights were freezed, then two LSTM layers were inserted between last two layers of architecture, Figure 4. These layers were trained and tested on new video data set. Training setting was same as with original weights training.

In Recognition-LSTM net.py is all essential functionality same as in Recognition-net.py, but directory to be analyzed is expected to have time sequence of images with filename containing their index in time. Option to choose CORAL weights is not available since weights for LSTM modified CORAL architecture have only one variant. Output is same but first 19 images lack age prediction, because for age inference there is requirement of 20 images.

5. Results

As CPU was used AMD Ryzen 5 2600 Six-Core 3.4 GHz Processor and for GPU NVIDIA GeForce GTX 1060 6GB. Inference speed of each CNN is in Table 1. Author of this paper used mostly PyTorch and had not installed driver requirements needed for GPU use for Keras and TensorFlow frameworks, therefore only CPU was used for gender and emotion inference. Average time of forward pass through LSTM layers was 1.642 ms (GPU).

Face detection AP (Area under curve AUC) was tested using both trained available RetinaFace architectures and weights in Table 2. Age prediction accuracy of CORAL network for each available weights

Table 1. CNN average inference speed, tested on FDDB data set [18] (1-4 faces). Inference speed for one image and LSTM modified version, used lightweight RetinaFace MobileNet-0.25 architecture

CNN network	CPU speed	GPU speed
RetinaFace Resnet-50	1.8668s	0.1539s
RetinaFace MobileNet-0.25	0.1399s	0.0243s
CORAL	0.1407s	0.0157s
CORAL with LSTM	1.1025s	0.0202s
Gender model	0.0048s	-
Emotion model	0.0032s	-
One image inference	0.4816s	0.1038s
Modified one image inference	1.3810s	0.1181s

Figure 6. Script output showing different age predictions for CORAL weights trained on different data sets (AFAD - top left, CACD - top right, MORPH - bottom left, UTK - bottom right)

was tested on UTK data set, Table 3. All weights were trained with different age ranges AFAD(15-40), CACD(14-62), MORPH(16-70), UTK(21-60), so to evaluate their performance, only images with persons in their age range intersection (21-40) were used for evaluation. Last compared weights were trained by author of this paper, using whole age (1-100) range of UTK data set. Metrics used to evaluate performance of age estimation were mean absolute error and root mean square error.

Table 4 shows comparison between CORAL model trained by author of paper on UTK data set and same model trained with additional LSTM layers. Since LSTM needs images in time sequence, new video data set was used for evaluation.

6. Conclusions

This paper describes basic scalable groundwork for further additional facial analysis.

Each network performance in terms of accuracy was satisfactory: RetinaFace predicts reliably as can be seen on face detection data set WIDER FACE on Table 2, gender network has 85.33% accuracy on IMDB celebrity data set[33], emotion network has accuracy 73.14% on fer2013 emotion data set [35]. Age estimation performance can be seen in Table 3, these values are not as good as authors had in their paper [5], this can be accredited to CORAL being trained on different sized crop than RetinaFace bounding box. This can be fine-tuned in the future. Overall the most underperforming prediction is age estimation but that is also the most difficult task. All data sets yielded good and bad results depending on person reflecting their data set nuances, this can be seen on Figure 6. CORAL Weights trained by author had yielded similar results to weights of original authors. Finally Table 4 and Figure 7, shows proof of improvement in CNN performance by introducing LSTM element, to harness temporal information. This result could be improved by enlarging video data set and changes to training.

Table 2. RetinaFace AP on validation data set WIDER FACE [39]

Architecture	Easy	Medium	Hard
Resnet-50	95.482%	94.046%	84.430%
MobileNet-0.25	90.708%	88.165%	73.827%

Table 3. CORAL network prediction error on section of UTK data set [40] in age range 21-40 (13147 images), for all training data sets options

Training data set	MAE	RMSE
AFAD	5.6004	7.0212
CACD	9.5130	11.9372
MORPH	7.4501	9.0592
UTK	6.3496	8.7732
UTK(1-100)	6.3516	8.7429

Table 4. Comparison between original and LSTM modified CORAL model, evaluated on new test video data set, 35 sequences - 35 unique individuals

CORAL Architecture	MAE	RMSE
Original	10.43	13.87
Modified	9.26	11.76

Figure 7. Script output showing different age predictions for original (left) and modified (right) CORAL architecture (trained on UTK(1-100), modified LSTM additionally trained on new video data set))

Inference speed of individual parts of image evaluation is on Table 1. Lightweight image analysis is around 0.1 s, but this highly depends on number of faces per image, so for more people in image prediction it could be problematic for real time use. Inference with CPU for more than 2 people is not usable and modified variant is not usable at all. This need for speed can be solved by further optimizing code between CNN passes or better hardware resources allocation. This can be done by not using neural frameworks by loading models as weights in ONNX format into performance-focused inference engine like ONNX Runtime [42], which can run atop of NVIDIA TensorRT Inference Accelerator [43].

Solution with LSTM, is now usable only for one person per image, because input needs time sequence of 20 bounding boxes of one person. So code should be updated to solve this issue for multiple persons.

This work in future expects to optimize inference speed and experiment with LSTM layer of CORAL network. Goal is to achieve live video feed inference performance with accuracy focused on age estimation.

Acknowledgements

I would like to thank my supervisor doc. Ing. Radim Kolář Ph.D. for his help.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. *Tensorflow: A system for large-scale machine learning*. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

[2] Afifi, M. and Abdelhamed, A., 2019. *AFIF4: deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces*. Journal of Visual Communication and Image representation, 62, pp.77-86.

[3] Antipov, G., Baccouche, M., Berrani, S.A. and Dugelay, J.L., 2017. *Effective training of convolutional neural networks for face-based gender and age prediction*. Pattern Recognition, 72, pp.15-26.

[4] Arriaga, O., Valdenegro-Toro, M. and Plöger, P., 2017. *Real-time convolutional neural networks for emotion and gender classification*. arXiv preprint arXiv:1710.07557..

[5] Cao, W., Mirjalili, V. and Raschka, S., 2019. *Consistent rank logits for ordinal regression with convolutional neural networks*. arXiv preprint arXiv:1901.07884.

[6] Chen, B.C., Chen, C.S. and Hsu, W.H., 2014, September. *Cross-age reference coding for age-invariant face recognition and retrieval*. In European conference on computer vision (pp. 768-783). Springer, Cham.

[7] Chen, J.C., Kumar, A., Ranjan, R., Patel, V.M., Alavi, A. and Chellappa, R., 2016, September. *A cascaded convolutional neural network for age estimation of unconstrained faces*. In 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS) (pp. 1-8). IEEE.

[8] S. Chen, C. Zhang, M. Dong, J. Le and M. Rao, "*Using Ranking-CNN for Age Estimation*," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 742-751. doi: 10.1109/CVPR.2017.86

[9] François Chollet, 2015 *Keras* GitHub, GitHub repository, [Online],

https://github.com/fchollet/keras, commit = 5bcac37

[10] Chu, W. and Keerthi, S.S., 2005, August. *New approaches to support vector ordinal regression*. In Proceedings of the 22nd international conference on Machine learning (pp. 145-152). ACM.

[11] Crammer, K. and Singer, Y., 2002. *Pranking with ranking*. In Advances in neural information processing systems (pp. 641-647).

[12] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. *Imagenet: A large-scale hierarchical image database*. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[13] Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I. and Zafeiriou, S., 2019. *RetinaFace: Single-stage Dense Face Localisation in the Wild*. arXiv preprint arXiv:1905.00641.

[14] Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., 2007. *The PASCAL visual object classes challenge 2007 (VOC2007) results*.

[15] Fan, Y., Lu, X., Li, D. and Liu, Y., 2016, October. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 445-450).

[16] FU, Cheng-Yang, et al. *Dssd: Deconvolutional single shot detector*. arXiv preprint arXiv:1701.06659, 2017.

[17] HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. *Long short-term memory*. Neural computation, 1997, 9.8: 1735-1780.

[18] Jain, V. and Learned-Miller, E., 2010. *Fddb: A benchmark for face detection in unconstrained settings*.

[19] Ko, B.C., 2018. *A brief review of facial emotion recognition based on visual information*. sensors, 18(2), p.401.

[20] Levi, G. and Hassner, T., 2015. *Age and gender classification using convolutional neural networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 34-42).

[21] Levi, G. and Hassner, T., 2015, November. *Emotion recognition in the wild via convolutional neural networks and mapped binary patterns*. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 503-510).

[22] Li, L. and Lin, H.T., 2007. *Ordinal regression by extended binary classification*. In Advances in neural information processing systems (pp. 865-872).

[23] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. *Microsoft coco: Common objects in context*. In European conference on computer vision (pp. 740-755). Springer, Cham.

[24] Lin TY, Goyal P, Girshick R, He K, Dollár P. *Focal loss for dense object detection*. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).

[25] Niu, Z., Zhou, M., Wang, L., Gao, X. and Hua, G., 2016. *Ordinal regression with multiple output cnn for age estimation*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4920-4928).

[26] PASZKE, A.; GROSS, S.; CHINTALA, S.; et al.: *PyTorch*. [online]. [Cit. 2019-05-10]. Retrieved from: https://pytorch.org

[27] Perez, C., Tapia, J., Estévez, P. and Held, C., 2012. Gender classification from face images using mutual information and feature fusion. International Journal of Optomechatronics, 6(1), pp.92-119.

[28] Polania, L., Fung, G. and Wang, D., 2019, January. *Ordinal Regression using Noisy Pairwise Comparisons for Body Mass Index Range Estimation*. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 782-790). IEEE.

[29] Ranjan, R., Sankaranarayanan, S., Castillo, C. and Chellappa, R., University of Maryland and College Park, 2019. *All-in-one convolutional neural network for face analysis*. U.S. Patent Application 16/340,859.

[30] REDMON, Joseph; FARHADI, Ali. *Yolov3: An incremental improvement*. arXiv preprint arXiv:1804.02767, 2018.

[31] REN, Shaoqing, et al. *Faster r-cnn: Towards real-time object detection with region proposal networks*. In: Advances in neural information processing systems. 2015. p. 91-99

[32] Ricanek, K. and Tesafaye, T., 2006, April. *Morph: A longitudinal image database of normal adult age-progression*. In 7th International Conference on Automatic Face and Gesture Recognition (FGR06) (pp. 341-345). IEEE.

[33] Rothe, R., Timofte, R. and Van Gool, L., 2018. *Deep expectation of real and apparent age from a single image without facial landmarks*. International Journal of Computer Vision, 126(2-4), pp.144-157.

[34] Shan, C., 2012. *Learning local binary patterns for gender classification on real-world face images*. Pattern recognition letters, 33(4), pp.431-437.

[35] Wolfram Research ,*The Facial Expression Recognition 2013 (FER-2013) Dataset*. From the Wolfram Data Repository (2018).

[36] Ullah, I., Hussain, M., Muhammad, G., Aboalsamh, H., Bebis, G. and Mirza, A.M., 2012, April. *Gender recognition from face images with local wild descriptor*. In 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 417-420). IEEE.

[37] Uricár, M., Timofte, R., Rothe, R., Matas, J. and Van Gool, L., 2016. *Structured output svm prediction of apparent age, gender and smile from deep features*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 25-33).

[38] XU, Yuanyuan, et al. *CenterFace: Joint Face Detection and Alignment Using Face as Point*. arXiv preprint arXiv:1911.03599, 2019.

[39] Yang, S., Luo, P., Loy, C.C. and Tang, X., 2016. *Wider face: A face detection benchmark*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5525-5533).

[40] Zhang, Z., Song, Y. and Qi, H., 2017. *Age progression/regression by conditional adversarial autoencoder*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5810-5818).

[41] ZHOU, X.,WANG, D., KRÄHENBÜHL, Philipp. *Objects as Points*. arXiv preprint arXiv:1904.07850, 2019.

[42] *ONNX Runtime: cross-platform, high performance scoring engine for ML models* [Online] at https://github.com/microsoft/onnxruntime

[43] *NVIDIA TensorRT™ - SDK for high-performance deep learning inference* [Online] at https://developer.nvidia.com/tensorrt

[44] *Facial recognition: top 7 trends (tech, vendors, markets, use cases and latest news)* [Online] at https://www.gemalto.com/govt/biometrics/facial-recognition

[45] *Google Scholar - Internet search engine for scientific papers* [Online] at https://scholar.google.com/

[46] *0-100 series, SoulPancace,Youtube* [Online] at https://www.youtube.com/playlist?list=PLzvRx-johoA-7DXw5EThWpQlVYH0hm1aC