POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

PAGE: advances in integrating pedestrian detection, age, and gender estimation

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Davide Foini

Advisor: Prof. Matteo Matteucci

Co-advisor: Simone Mentasti

Academic year: 2022-2023

1. Introduction

In the past few years, pedestrian analysis, notably pedestrian detection, has emerged as a significant aspect within object detection, finding utility in autonomous driving, surveillance, and tracking systems. Additionally, the subset of Pedestrian Attribute Recognition (PAR), focusing on age inference from face images and gender identification from full-body representations, has gained considerable attention.
The majority of these applications are tailored for mobile device use, primarily because of the specific settings needed and privacy concerns. Nonetheless, current research often neglects the hardware limitations associated with these aspects [1]. Additionally, current age estimation models often exhibit performance deficiencies when tested across varying datasets, particularly suffering from unbalanced representations across different age ranges [8]. In addition to this problem, age estimation suffers from intra-class variation. Finally, age estimation methods can adopt either a regression or classification approach [2].
Having considered these aspects, this work provides two contributions. The first one is PAGE (Pedestrian Age and Gender Estimation), a framework merging pedestrian detection, age regression, and gender classification. Its purpose is to operate seamlessly online and on mobile devices, combining tasks that are usually tackled independently. The second contribution is to propose a model that performs age classification based on age groups focusing on the performance on different age values and not only on the average one.

2. The PAGE Framework

The pipeline comprises three primary modules: YOLOX-nano and the Head Analysis and Body Analysis modules. Initially, YOLOX [3] processes the input image to extract detections, subsequently filtered to derive body and head images. These sub-images undergo preprocessing and batching before analysis by the Head and Body Analysis modules. The resulting labels are then integrated into the detected bounding boxes, culminating in the head-body association. The resulting image showcases detection boxes alongside corresponding gender and age labels. An overview of the framework is available in Figure 1. YOLOX has been selected among the detectors available due to its good balance between accuracy and speed, besides

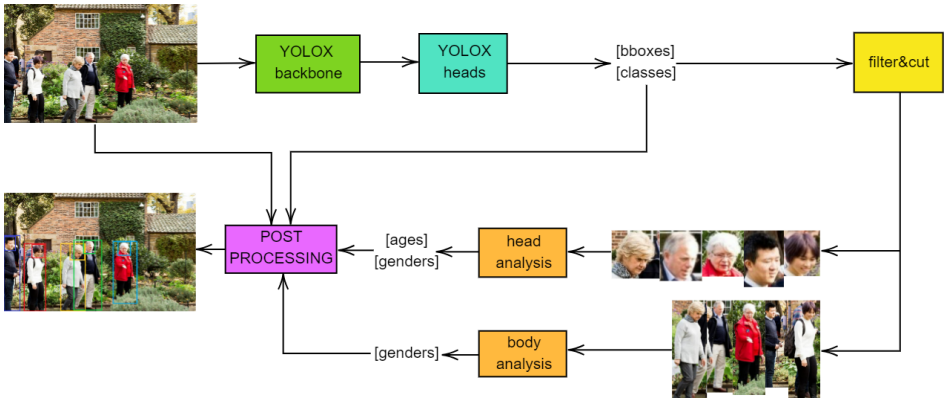


Figure 1: An overview of the PAGE framework.

the fact that full code and extensive documentation are made available by the authors. Moreover, the YOLOX-nano version has been considered because it has been specially designed to run on mobile devices. YOLOX utilizes the YOLOv5 Focus module for channel optimization and the CSPDarkNet feature extraction network from YOLOv4 as its backbone. The YOLOX neck uses the PAFPN structure to blend multi-scale features seamlessly. Following processing through the backbone and neck networks, the input image is segmented into three scales for predicting large, medium, and small objects. The next step is to feed the features from these scales into the decoupled detection heads. Each head is responsible for finding the bounding boxes and the respective object class. Since it is common to detect more bounding boxes for the same object, NMS is performed to keep only the best box for the object.

The preprocessing functions are necessary to resize the images to the expected size by the Head and Body Analysis modules. For head images the expected size is 80 x 50 and 250 x 100 for the body images.

The Head Analysis module is responsible for taking as input a head image and estimating the gender and age of the subject. The main idea is to exploit the features extracted by CSPDarknet (YOLOX backbone) to perform the predictions. Two different networks are employed: one for age regression and one for gender classification. Once the labels are obtained they are sent to the post-processing stage.

The loss used for age regression is the L1 loss, which measures the MAE (Mean Average Error) between the output of the network and the target values. For gender classification, the chosen loss is the BCE with logits loss, which combines a *Sigmoid* function with the Binary Cross-Entropy (BCE) loss.

The Body Analysis module is similar to the Head Analysis one, but it only estimates the gender given the full body picture. It uses the same backbone and then a network tailored to perform gender classification. The estimated gender label is then post-processed with the labels generated by the Head Analysis module, the detection boxes and the input image. The training loss is always the BCE with logits.

After obtaining the age and gender from the head picture and the gender from the body image, the bounding boxes need to be associated to link the head and body. This problem is solved by analysing the head boxes and for each of them computing the Interception over Head (IoH) with the body boxes, and then associating it with the one with the maximum score. The IoH is obtained with the following formula:

$$IoH(h,b) = \frac{A(h \cap b)}{A(h)} \tag{1}$$

where h and b are the head and body bounding boxes, $h \cap b$ their intersection and the function $A(x)$ computes the area of the given box. It follows that $IoH \in [0,1]$, where $IoH = 0$ when there is no intersection and $IoH = 1$ when the head box is fully included in the body box. The final result is composed of different bounding boxes, each with the label class, the gender class, and the age class if it belongs to the head class.

2.1. Experiments

YOLOX, originally trained on the COCO dataset comprising 80 different object classes, lacked representation for detecting human heads. To address this limitation, we fine-tuned YOLOX on the CrowdHuman Dataset for 300 epochs.
The datasets used to train the gender classification model from facial images are AFAD (Asian Face Age Dataset), AgeDB and UTKFace. The utilized model leverages YOLOX's initial feature level, followed by two additional convolutional blocks and a classifier. These added convolutional blocks aim to enable the model to acquire new features, compensating for the backbone's original training focused on object detection, not classification. Importantly, the YOLOX backbone is intended for use without fine-tuning, meaning that it is not trained to perform gender and age estimation but only for object detection. To perform age regression the dataset chosen was the one obtained keeping a maximum of 1500 images for each age value. This dataset is later described in Section 3 with the data augmentation used. Different models have been tested, the first is similar to the one used for gender classification that exploits some of the features extracted by CSPDarknet53. The other two models have ResNet50 and MobileNet V2 as backbones respectively.
To train the gender classification model from full-body images the PA-100K and PETA datasets were chosen. Following the same approach used for age regression from facial images, two different models have been tested to perform gender regression from full-body images. The first model always exploits the backbone of YOLOX-nano, followed by convolutional blocks and then a classifier, while the second is a custom CNN.
The last experiment tested the actual performance of the pipeline in a real-time scenario. To have a better understanding of the weight of each different operation, we decided to report the different modules and their respective average time. The experiment has been carried out in a lab environment using a docker container. The hardware used is composed of an NVIDIA GeForce GTX 1080 Ti GPU with an Intel Xeon E5-2630 v4 CPU.

2.2. Results

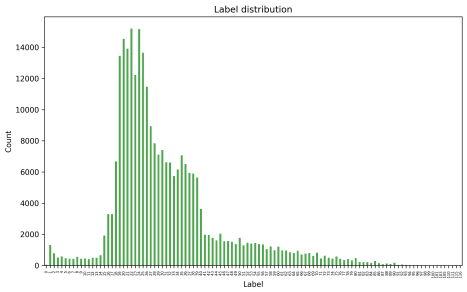
After the fine-tuning process, the model obtained an average precision (AP) of 21 for body boxes and 17 for head boxes, whereas the YOLOX-nano pretrained model had an mAP (mean Average Precision) of 25.8. From these examples, it is possible to note how in some settings where the subjects are not too many and quite close to the camera the detections are pretty accurate. However, when the subjects are at a higher distance or there is a crowd the detector can fail. The gender regression model achieved a test set accuracy of 92.56%. Regarding age estimation, the model with CSPDarknet as its backbone achieved a Mean Absolute Error (MAE) of 7.89. Comparatively, models employing ResNet50 and MobileNet V2 achieved MAEs of 7.46 and 8.17, respectively. While these models demonstrate commendable performance in age estimation, accurately predicting age remains challenging due to limited model sizes, augmented data, and the low resolution of the images.
The body analysis models achieved gender estimation accuracies of 72.63% and 72.87% when working with full-body images. This decrease in accuracy, compared to head images, is attributed to the head being the most distinctive area for gender prediction, while the rest of the body tends to exhibit greater similarity among individuals. Consequently, this similarity results in fewer discriminative features, impacting the gender classifier's accuracy in body analysis.
In Table 1 we reported the average execution time and the resulting FPS for each module. *Detection* involves utilizing the YOLOX model to identify heads and bodies. *Association* aims to pair each head with its corresponding body. *Filter&Cut* involves extracting head and body images from the input image using YOLOX detections. *Head and Body Analysis* estimate age and gender from facial and full body images. *Other* operations encompass additional tasks such as plotting bounding boxes and labels on the final image. It is possible to note how on average the most expensive phase is the detection one, followed by the other operations, the head analysis, the body analysis and the filtering and cutting task. The least expensive operation is the box association. As expected, the head analysis module takes more time than the body analy-

Task	AVG Time (s)	FPS
Detection	0.1555	6.43
Association	0.006	166.67
Filter&Cut	0.0409	24.45
Head Analysis	0.0813	12.3
Body Analysis	0.0627	15.95
Other	0.0824	12.14
Total	**0.4287**	**2.33**

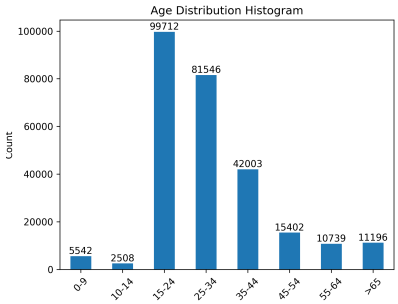
Table 1: Execution times for different runs and average values.

Dataset	Total Images	Age Range
FGNET	1 002	0 - 69
UTKFace	23 708	1 - 116
AgeDB	16 488	1 - 101
APPA-real	7 591	1 - 100
KANFace	41 036	0 - 98
AAFD	13 322	2 - 80
AFAD	165 501	15 - 72
TOTAL	**268 648**	**0 - 116**

Table 2: The datasets considered and their composition.



(a) Tot: 268 648 - Min: 0 - Max: 116



(b)

Figure 2: Total age and group distribution.

sis one, having to compute both gender and age labels, but not double the time since body images have higher resolution (80 x 50 and 250 x 100). The achieved operating frequency of 2.33 FPS remains satisfactory, especially considering the scenario of walking pedestrians. This assessment holds, given the approach of combining four distinct models and executing operations on images which can be particularly resource-intensive. The detection tasks can infer at 6.43 FPS, while Tiny-YOLOv3 runs at 4.14 FPS [6]. Greco et al. [4] proposed a real-time gender recognition model from face images that can operate at 5 FPS, while our Head Analysis module that performs both age and gender estimation does so at 12.3 FPS. The resulting accuracy when performing gender estimation is 94.99%, similar to the one obtained by the Head Analysis module of 92.56%. An optimal solution is presented in [7], where a real-time single-shot multi-face gender detector based on a CNN can infer at 83 FPS. This suggests how a model where only one pass in the network is necessary can significantly increase the operating frequency.

3. Age Groups Classification

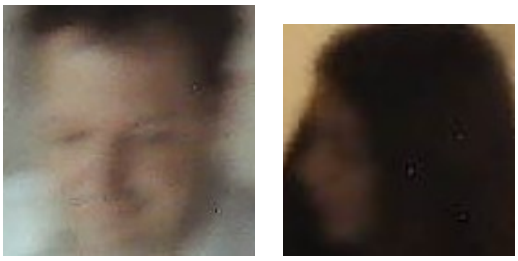
The second part of this thesis focused on the development of a model that could classify a facial image into one of the eight age groups, which are 0-9, 10-14, 15-24, 25-34, 35-54, 55-64 and 65+. The datasets considered are FGNET, UTKFace, AgeDB, APPA-real, KANFace, AAFD (All-Age-Faces Dataset), and AFAD (Asian Face Age Dataset). In Table 2 the dataset sizes and age ranges are reported. The difference in size and age ranges between the datasets is quite noticeable. The main issue is the class imbalance: all the datasets tend to have a higher concentration of samples in specific ranges. When the datasets are combined, this problem is worsened, as possible to note in Figure 2. Most of the samples have labels in the range 18-35 and therefore 37.11% in the 15-24 class, 30.35% in the 25-34 range and 15.63% within the 35-44 range. Using this data to train a model would have the obvious consequence of overfitting the data inside those ranges and having very inaccurate predictions when the true age belongs to the other five range classes. To try to mitigate this issue we decided to balance the data keeping only 5000 samples for each range. To do so we needed more samples in the 10-14 range since only 2506 elements are provided in this class. To obtain

more images we opted to create new ones starting from the ones already available and applying some methods of light data augmentation. Other ways of balancing the data have been to keep a fixed maximum number of samples per age value, for example, 1000, 1500 or 2000.
A strong data augmentation was employed during training. The modifications applied include a brightness change from 50% to 150%, a channel shift with a range of 128, a maximum rotation of 30° and a zoom ranging from 0.75 to 1.0. This choice stems from the datasets' original images captured under ideal conditions with high quality and optimal lighting. However, this work specifically addresses surveillance camera scenarios, which often present challenges such as low resolution and varying illumination, differing from the ideal conditions in the datasets. State-of-the-art models usually consider input images to have at least a medium size. Analysing the use case scenario of a HD surveillance camera we noticed how detected boxes in most cases have a small size, ranging from 50 x 50 to 100 x 100, therefore we selected 75 x 75 as a reasonable input size. We investigated how much this reduction in the amount of features extracted can affect the prediction performance of a model. It was tested on all the datasets and the one composed of a maximum of 5000 images per range and a maximum of 1500 samples per age value. It is not surprising to find that as the input size increases, the mean absolute error (MAE) of the model decreases across all datasets. The minimum MAE decrease of 7% is observed on the AFAD dataset, while the maximum MAE improvement of 26% is seen on the dataset with a maximum of 5000 samples per age range. On average, the MAE decreases by 17%. When analysing the different age ranges, with a higher resolution the model improves on all the range classes besides the last one. The highest improvement is in the second class with a +16% in accuracy and the accuracy in the last range is decreased by 3%. On average the accuracy improves by 8.36%.
When comparing a regressor with a classifier with one class for each range, we observed the best performance is given by the regressor, while the classifier is subject to completely wrong classifications, such as classifying images in the 25-34 range as a 0-9 class. The remapping process of the output of the regressor is carried out via a lambda layer that, given the estimated age, converts it into a vector as if it would have been the output of a standard classifier. The lambda layer applies a sequence of normal distribution, one for each age class so that the final vector has in each position the value of the corresponding distribution for that age. Combining individuals of different ages into a single group can make it more challenging to identify shared characteristics, rather than simplifying the process. Another major downside of a range classifier is that in case the desired age ranges are changed, the model has to be retrained, while with the regressor it is enough to change the lambda layer and the distributions at the end of the network. We also compared a regressor with a classifier for each age value, but we did not find a consistent difference between the two.
MobileNet V2 [5] has been chosen as the baseline network because it is a very lightweight model and for its popularity in mobile applications where the computational power is limited. We have observed that training with 1000 images per age value instead of 5000 per range can lead to improved performance in specific ranges. For instance, in the age range of 15-24, a noticeable improvement of +10% has been registered, and in the age group of 65+, there is an improvement of +16%. However, we have noticed a decline in performance in the age ranges of 25-34 and 55-64. The model with 5000 images per range has a MAE of 6.78, while the other has a MAE of 6.85, indicating similar overall performance.
When using the datasets with 1000, 1500 and 2000 images per age value, the MAEs recorded are 5.86, 5.5, and 5.76, respectively. Notably, the most favourable outcomes, with the highest accuracy in five out of eight range classes, are observed when utilizing a maximum of 1500 samples per age value. These results are detailed in Table 3. A different test performed released the constraint of using MobileNet as a baseline and it tried to investigate if employing a larger model would improve the prediction performance. The different backbones tested besides MobileNet V2 are VGG16 and VGG19, ConvNeXt base and large, EfficientNet V2 Large, InceptionResnet V2 and ResNet50 V2, ResNet101 V2 and ResNet152 V2. Larger

Range	1000	1500	2000
0-9	84	**87**	84
10-14	52	**57**	37
15-24	**63**	58	60
25-34	53	**58**	56
35-44	33	**43**	39
45-54	51	**53**	52
55-64	**51**	46	44
65+	**71**	65	66
MAE	5.86	**5.5**	5.76

Table 3: Results when using datasets with a maximum of 1000, 1500 or 2000 per age value. The best results are reported in bold.

models have on average a better accuracy, except for the 25-34 range where MobileNet has the best score. In particular, for some classes, the performance improvement can reach significant values for the baseline, like +22% in the 10-14 range, +18% for the 15-24 class and +12% for 65+. Of course, the model size has to be taken into consideration in terms of memory occupancy, training and inference times, but having larger models can result in better accuracy. When compared to state-of-the-art models the results are worse, with an average increase of 3 in the MAE. This is comprehensible given that our models are trained with all the datasets mixed, so they are not prone to overfit on the data of every different dataset, and they handle images with lower resolution (75 x 75), as already mentioned in this section.
The last series of tests carried out involved new data, acquired in a challenging scenario with a low-resolution camera in an office room. The samples are extracted from video sequences for a total of 267 images. This new dataset is very limited in size and age values. Another downside is that, being taken from a video sequence, most of the samples are taken from consecutive frames, therefore the images are very similar. For these reasons and to simulate a real-case scenario, the dataset has been used only as a test set. In Figure 3 some of the samples are shown. We tested models with MobileNet V2 or VGG19 as the backbone and trained on the datasets with a maximum of 1000, 1500 or 2000 samples per age value. The VGG 19 backbone trained with a maximum of 2000 samples per age value stands out as the best model, achieving 31% accuracy



(a) Age: 48 (139 x 139) (b) Age: 34 (86 x 86)

Figure 3: Examples of samples from the challenging dataset.

in the 25-34 range and 30% in the 35-44 class. Models utilizing MobileNet V2 as a backbone generally exhibit improved performance with an increase in the number of samples, except for the 45-54 age group. Increasing samples tends to enhance average accuracy.

4. Conclusions

Fine-tuning the detector on the CrowdHuman dataset improved head detection, yet challenges persisted with high subject density and distance affecting accurate detection. Notably, the limited model size hindered precise age and gender estimation due to low-resolution head and body images.
System operation analysis revealed detection as the most resource-intensive task followed by box and label projection. Future work could focus on an integrated end-to-end model or explore alternative methods like attention mechanisms for enhanced prediction performance. Robust association methods in crowded scenarios remain an area for further research.
Additionally, efforts were made to develop an age estimation model classifying facial images into eight age ranges. Experimentation with common datasets revealed insights, favouring a regressor-based approach over direct classification for more practical outcomes.
Key considerations include the significant impact of image resolution on accuracy and the trade-offs associated with model size. When new and challenging data is tested, the models exhibit expectedly lower accuracy, but performance can be improved by using more samples. Future research could explore new CNN models, use attention mechanisms, or leverage different feature types.

References

[1] Farhat Abbas, Mussarat Yasmin, Muhammad Fayyaz, and Usman Asim. Vit-pgc: vision transformer for pedestrian gender classification on small-size dataset. *Pattern Analysis and Applications*, pages 1–15, 2023.

[2] Arwa S Al-Shannaq and Lamiaa A Elrefaei. Comprehensive analysis of the literature for age estimation from facial images. *IEEE Access*, 7:93229–93249, 2019.

[3] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[4] Antonio Greco, Alessia Saggese, and Mario Vento. Digital signage by real-time gender recognition from face images. In *2020 IEEE International Workshop on Metrology for Industry 4.0 and IoT*, pages 309–313, 2020.

[5] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[6] Hyunduk Kim, Myoung Kyu Sohn, and Sang Heon Lee. Development of a Real-Time Automatic Passenger Counting System using Head Detection Based on Deep Learning. *Journal of Information Processing Systems*, 18(3):428–442, 2022.

[7] Tak Wai Shen, Dongpeng Wang, Kayton Wai Keung Cheung, Man Chi Chan, King Hung Chiu, and Yiu Kei Li. A real-time single-shot multi-face detection, landmark localization, and gender classification. In *Proceedings of the 2021 3rd International Conference on Image Processing and Machine Vision*, pages 1–4, 2021.

[8] Beichen Zhang and Yue Bao. Cross-Dataset Learning for Age Estimation. *IEEE Access*, 10:24048–24055, 2022.