A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

Open source soft biometric feature extraction: a Benchmark

Tizian Rettig[1]

Abstract: The field of soft biometrics has become more and more important over the years and has been tackled largely by Machine Learning approaches. Though these types of approaches promise huge success in accuracy, the comparability of such models is questionable at best. To provide a more even playing field, this paper discusses a general benchmark for a wide range of classifiers. A set of definitions for commonly used terms and measures is provided and reference results are presented. It is shown how the benchmark avoids over simplification while being able to break down the performance into easily digestible values. The main focus of this paper is the evaluation of open source models, but any classifier would suffice.

Keywords: Soft biometrics, Benchmark, Age prediction, CNN, RoR-34, Framework, Open source, Machine learning

1 Introduction

The field of biometrics has matured to a point in which identification of subjects based on unique features is well understood. It is possible to recognize the identity of a subject based on their iris pattern, finger prints or facial structure. All of the afforementioned approaches have been tested and used in a commercial context. However, as these approaches are designed to work in highly controlled environments, their accuracy drastically decreases when presented with low quality samples. Looking at facial recognition in particular, it is shown that factors like face angle or lighting conditions reduces accuracy drastically [Br98]. To be able to achieve similar performance for images taken ”in the wild” (meaning in uncontrolled environments like surveillance camera footage or group pictures), the scientific community has started to look into secondary features that could provide valuable information for the classification of subjects. These so called soft biometric features range from temporary traits like cloth to more per permanent features like hair color or the gait of a person. Even broader traits like age or gender are often targeted in modern research papers. As the landscape of this field has brought many novel approaches and performance claims, this paper aims to propose a benchmark applicable to soft biometric feature detectors in general. This is done to provide a objective look at the current state of the art and a baseline for further research.

2 Soft Biometrics

To base this paper on a solid foundation, a concrete understanding of the Terminology used is essential. Therefore, available literature on biometrics in general and soft biomet-

[1] Hochschule Darmstadt, Schöfferstraße 3, 64295 Darmstadt, tizian.rettig@stud.h-da.de

rics in particular were found. The most noteworthy of which were two surveys on the state of the art from 2016: ”What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics”[DER16] and ”On Soft Biometrics”[Ni15] as well as a very recent survey, ”Soft biometrics: a survey”[Ha21] from 2021. In which the following definition is proposed:

> Soft Biometric traits are physical, behavioral, or material accessories, which are associated with an individual, and which can be useful for recognizing an individual. These attributes are typically gleaned from primary biometric data, are classifiable in pre-defined human understandable categories, and can be extracted in an automated manner.

Besides a useful definition of the term soft biometric trait, these three surveys present open questions, further research topics and lists of available resources related to the field. These resources were heavily used in the research for this paper.

3 Benchmark design

As the benchmark aims to provide a standardized look at the objective performance of a given detector, it provides open source interfaces to extend the existing data sources and detectors. In the beginning it will only address a single soft biometric feature (age) and present a foundation for the future. To be able to provide meaningful results broadening the scope of the paper too much, only one dataset and two detectors will be looked at. This first iteration will specifically look at the CNN based approach shown in Tal Hassner and Gil Levi’s 2015 work[3] [LH15] as well as the RoR based approach presented in ”Age Group and Gender Estimation in the Wild with Deep RoR Architecture”[4] published in 2017 [Zh17]. The dataset implemented for this paper is the Adience dataset introduced in 2014 [EEH14].

3.1 Feature selection

When choosing the feature that should act as the reference for future implementations a few key aspects have been taken into a count.

1. The feature should be commonly discussed in state of the art research of the field.
2. It should be available in many datasets
3. It should not be too sensitive as to avoid heavy NDAs on datasets or controversy
4. The problem it presents should be relevant for the foreseeable future

[3] Source code: https://github.com/GilLevi/AgeGenderDeepLearning

[4] Source code: https://github.com/ShreyanshJoshi/Facial-Demographics-using-CNN

To accommodate for all the points mentioned above, the extensive list presented by [Ha21] was looked at. In the survey they present over 100 soft biometric features that have occurred in recent research. The list is split into several categories namely temporary modalities, permanent modalities related to the face or head and permanent modalities related to the body. Further more, so called global traits are discussed. The traits mentioned in this category are: age, gender and ethnicity.
From these categories the group of temporary modalities was discarded as a candidate early on. Such features might pose too complex of a problem for the task at hand. They also might be more niche in nature as it would require rather specific circumstances for clothing or the like to be relevant in subject recognition. Taking a closer look at the modalities related to the body it becomes apparent that most traits in this category also favor certain circumstances. For example attributes like ”Shoulder shape” or ”Hip width” seem rather unique. Looking at more common (based on the number of citations mentioned in [Ha21]) traits related to the general body we find features like ”Weight” or ”Height”. Both of which are harder to collect than images of faces, for which many data sets already exist and are often shared on social media and the likes. Taking a look at the remaining large category of face modalities, we find some very loosely defined attributes like ”Face type”. Focusing instead on more clearly defined values like ”Nose width”, ”Hair color” or ”Eyebrow length” we still get little inconsistencies or ambiguities that need to be answered before addressing a given feature. Where would the nose width be measured? How would the hair color be discretized? What if somebody shaves their eyebrows? To avoid all of these, the only soft biometric features worth using for this benchmark are global traits. Based on the benchmarks presented by the website paperswithcode.com[5] gender prediction has come a long way while age estimation still poses some challenges. This makes age a great choice for being a reference feature.

3.1.1 Ethnicity

A attribute that should also be mentioned is ethnicity. Ethnicity has become an emotionally charged topic over the past few years. Therefore it should be treated with utmost care. Taking a look at some recent works in the field that predicted ethnicity as one of its features, it feels like the buckets/labels used as possible values missed the mark in multiple aspects. All of which come down to the issue of grouping individuals together based on ambiguous or miss guided terminology [Ma18]. To avoid controversy and to ensure the topic is dealt with properly, ethnicity is avoided as a candidate in this benchmark. More time should be invested before adding this value.

3.2 Implementation

The benchmark tool proposed in this is a slim python module that consists of three components that should be inherited to connect more datasets and detectors to the benchmark. These are:

[5] https://paperswithcode.com/task/age-and-gender-classification

1. The DataSet class - it represents the actual dataset files stored on disk and provides access to them through chunked pandas.DataFrames
2. The Detector class - this class is a slim wrapper that is called by the core component to execute classifications using the algorithm implemented in it
3. The FieldTranslator class - its main purpose is to append the groundtruth fields found in a given data set to the classification result in a manner that it is easily comparable to the predictions made by the detector class.

Using these three components, it should be possible to interface with the core component of the system, which is able to evaluate the results generated with the detector.

The intention of this is that future research groups that want to test a given model are able to use a broad range of data sets with little implementation overhead. New datasets could also be introduced regularly until most of the common datasets are available and the evaluation of new models can be done on a large number of datasets. This provides information about generalization of the given model and the relative performance compared to other models. It also avoids the issue of datasets being unintentionally biased as many sources will lessen the effect of such factors. (The source code is available under the Github mentions in the foot notes [6])

3.3 Metrics

When talking about benchmarks one of, if not the most important aspect is the evaluation of the generated results. The core component of the benchmarking tool saves the result in series of CSV files. Each of which represent one chunk of the dataset and its corresponding predictions generated with the given detector. A sample snipped of one such chunk is shown in the following[7]:

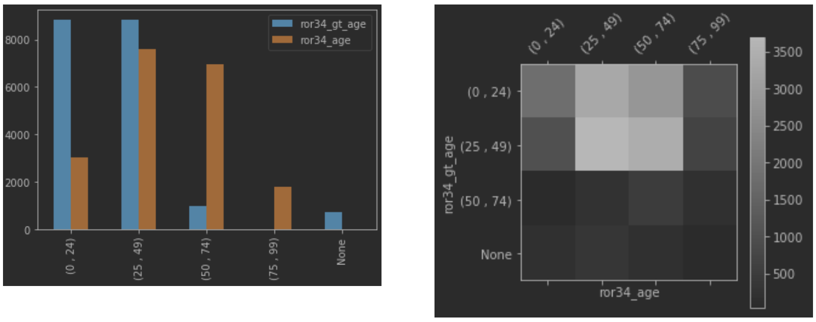
```
,age,img_path,LeviCNN_age
0,"(25, 32)",/mnt/f/.../10399646885_67c7d20df9_o.jpg,"(60, 100)"
1,"(25, 32)",/mnt/f/.../10424815813_e94629b1ec_o.jpg,"(25, 32)"
2,"(25, 32)",/mnt/f/.../10437979845_5985be4b26_o.jpg,"(25, 32)"
3,"(25, 32)",/mnt/f/.../10437979845_5985be4b26_o.jpg,"(4, 6)"
```

Having this representation, it is possible to analyse the data using a jupyter notebook or the provided methods of the core component. The following values should be generated if the result is meant to conform to the rules of the benchmark presented in this paper.

[6] Github source of the initial implementation as well as supplementary materials: ”https://github.com/Saphs/sbf$_{e}xtractor_{b}enchmark''$

[7] Taken from the results of the LeviHassner CNN approach [LH15]

Fig. 1: Example results from RoR-34 [Zh17] on Adience [EEH14]



(a) age spread

(b) confusion matrix

1. A histogram comparing the spread of all possible values in the ground truth against all predicted values. As an example, see figure 1(a). This is plotted to inspect whether or not the predicted values follow a random distribution or any biases towards certain buckets are visible. Further more, a confusion matrix should be plotted. This will display similar information, but on a value by value basis. Such a matrix can be seen in figure 1(b).

2. Accuracy, which is split into four specific values. First of all ”Accuracy” it self defined as:

$$Acc = \frac{correctPredictions}{totalPredictions}$$

This provides a basic understanding on how likely it is to receive the expected out come. This value is typically provided but can hide some technical details which is why we introduce ”Baseline Accuracy” defined as:

$$Acc_0 = \frac{1}{numberOfBuckets}$$

With this we provide an expectation on how a uniformly distributed random number generator (later called RNG) would perform in the same situation. If, for example, a detector performs with an $Acc = 55\%$ and an $Acc_0 = 50\%$ due to the detector only having two buckets to choose from it essentially performs only 5% better than a coin flip. Which brings us to the third value. ”Accuracy over random chance” defined as:

$$Acc' = Acc - Acc_0$$

It presents the absolute percentage value the detector performs better than a RNG. The forth value related to accuracy (”Accuracy score”) introduces a non-technical scoring system that presents the accuracy of the detector as an easily comparable number. It is defined as follows:

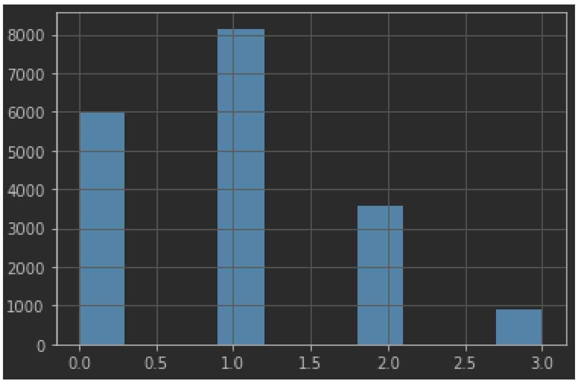
$$Acc_{norm} = \frac{Acc'}{1 - Acc_0}$$

$$Acc_{score} = \lceil Acc_{norm} * 1000 \rceil$$

The metric describes the accuracy of the detector on a scale from 0 to 1000 based on its normalized accuracy over the randomized approach. Therefore a low score will indicate that the detector is barely more accurate than guessing and vice versa.

3. A histogram that shows the distribution of error values. An example for this can be seen below in figure 2. Its intention is to better asses if the mean error is effected heavily by outliers.

Fig. 2: error spread from RoR-34 [Zh17] on Adience [EEH14]



4. Mean absolute error - Another common way of comparing detectors uses the mean absolute error (”MAE”). It provides an intuitive metric that represents the magnitude of the error that is to be expected and is defined as:

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

However, it fails to provide a meaningful metric as soon as differing bucket count or sizes are introduced. Therefore a normalized MAE is defined like this:

$$MAE_{norm} = \frac{MAE}{numberOfBuckets - 1}$$

with $numberOfBuckets - 1$ representing the maximum possible error. Finally another scoring value is introduced. The MAE_{score} value is based on MAE_{norm} and provides a non-technical look at the expected error per prediction. It is defined as:

$$MAE_{score} = \lceil (1 - MAE_{norm}) * 1000 \rceil$$

In essence the scoring value is the inverse of the MAE_{norm} and scaled into the range between 0 and 1000. Therefore a low score indicates a high MAE.

5. The final value presented here, is the Combined score. This represents the highest abstraction level and is a non-technical measure to compare models at a glance. It is defined as follows:

$$score = Acc_{score} + MAE_{score}$$

and ranges from 0 to 2000. With 2000 being a model that has a accuracy of 100% and a mean absolute error of 0.

4 Experimental results

In this section the benchmark results for the LeviHassner CNN and the RoR-34 approach are discussed. It should be noted that the RoR-34 approach was not trained using the Adience dataset while the LeviHassner CNN was. This is very noticeable as it seams that the RoR-34 does not generalize to this dataset too well. These types of results show the benefit of running such a broad benchmark that may present novel images to a pre-trained model resulting in a real world test for generalization. Having said that, the UTKFace dataset[8] that was used to train the RoR-34 approach is freely available. It was excluded from the experiments due to the lack of time and the small scope of this paper. Further time should be invested to give a more fair comparison of the two models. As it stands, the results are heavily tilted in favor of the LeviHassner CNN and should not be interpreted to discredit the results achieved by RoR-34.

4.1 LeviHassner CNN

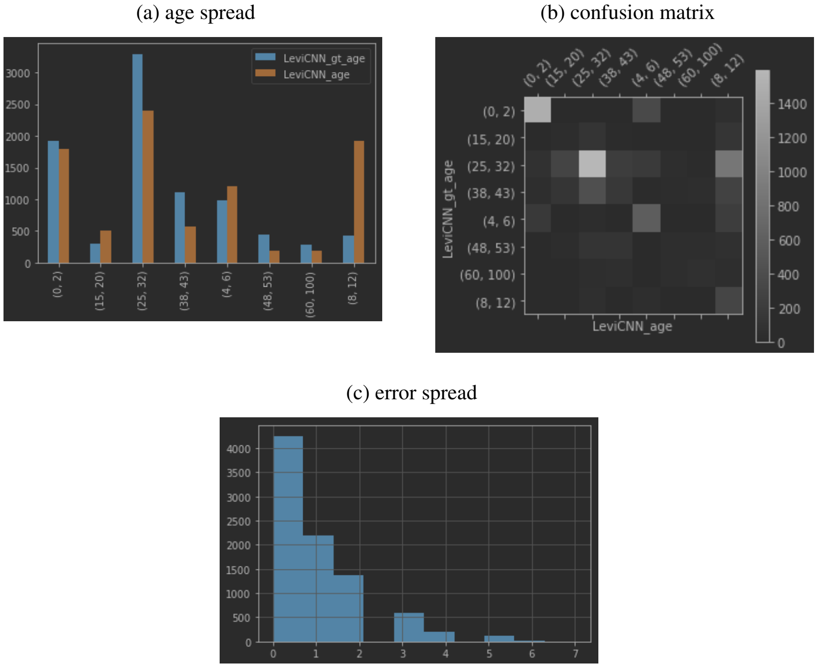
In figure 3 we can see the plots required by the benchmark description given earlier. We see a rather successful attempt at predicting the ages provided in the Adience dataset. This however is to be expected because the model was presented here was trained on the adience dataset. It should also be noted that the claimed accuracy of 44 - 50%[9] stated on the benchmark leaderboard could be reproduced. Taking a look at table 1 we can see that its score came to be 1265 with a Acc_{score} of 399 and a MAE_{score} of 866. This shows that the model is rather successful in predicting a rough estimate but fails to actually correctly predicting the age often. Looking at the plots in figure 3 again, more specifically the confusion matrix (b), we can see that the model has some issues predicting the age of 25 to 32 year olds. It seem they are mistaken to be between 8 and 12 rather often.

Tab. 1: Numeric results for the LeviHassner CNN [LH15]

Dataset	$score$	Acc_{score}	MAE_{score}	$numberOfBuckets$
Adience	1265	399	866	8

Acc	Acc_0	Acc'	MAE	MAE_{norm}
0.4743	0.125	0.3493	0.9403	0.1343

Fig. 3: Results from LeviHassner CNN [LH15] on Adience [EEH14]



(a) age spread

(b) confusion matrix

(c) error spread

4.2 RoR-34

Switching to the second result set build using the proposed benchmark, we see a less promising result. This is expected as the model was not trained on this data set and seems to fail at generalization. In this example the usefulness of the age spread plot (figure 4(a)) becomes apparent. The predicted values (though few in numbers) seem to trace a Gaussian distribution. This could indicate that the model does not perform well on the newly presented data. We see evidence for this in both the confusion matrix and the error spread as well. The poor performance is also reflected in the table 2. In which we see a Acc_{score} of 79 and a Acc' of 6%, making the model barley better than randomly choosing an option. Assuming the results are no statistical fluke, which would be possible for such a low value.

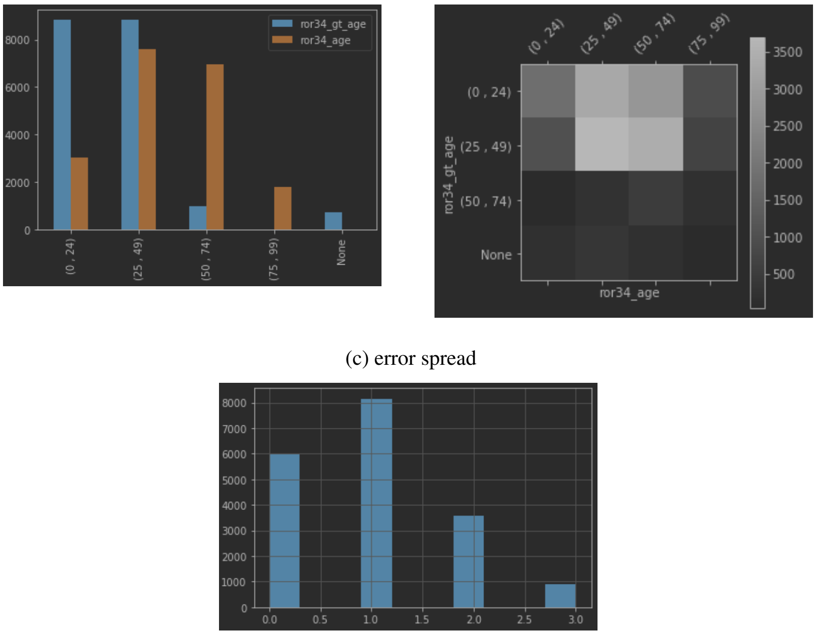
4.3 Comparison

The results of both models show the drastic impact missing generalization has on the performance of a network. While the age spread of the LeviHassner CNN clearly resambles

[8] Available here: https://susanqq.github.io/UTKFace/

[9] The Multiple versions of the LeviHassner CNN can be found on the leaderboard, see https://paperswithcode.com/sota/age-and-gender-classification-on-adience-age

Fig. 4: Results from RoR-34 approach [Zh17] on Adience [EEH14]



(a) age spread

(b) confusion matrix

(c) error spread

Tab. 2: Numeric results for the RoR-34 approach [LH15]

Dataset	*score*	Acc_{score}	MAE_{score}	*numberOfBuckets*
Adience	757	79	678	4

Acc	Acc_0	Acc'	*MAE*	MAE_{norm}
0.3094	0.25	0.0594	0.9661	0.3220

the spread of the groundtruth, the RoR-34 approach does the opposite. A similar picture is painted when looking at the error spread which also shows drastic differences between the model performances.

5 Conclusion

In this paper a novel benchmarking approach was presented. It was shown that a non-technical representation to ease comparisons and a in depth look into the performance of a given model could be achieved this way. The approach is meant to grow in significants as datasets are added and the resulting conclusions for each model become more meaningful. A general idea on how to implement new datasets and connect existing models to the application was provided. Future research could mature the application to a point at which it is simple to test a newly developed model against known datasets in a plug and play fashion. Further time needs to refine both the implementation and analysis.

6 Outlook

As the work done in this paper is only the foundation for future work, its use largely depends on the number of datasets connected to in the future. Therefore future research could invest time into adding a multitude of datasets. It would also be worth discussing the value of evaluating models on the pool of datasets available with one single combined result or on a dataset by dataset basis. This paper opens many possible research questions that could be addressed in the future.

References

[Br98] Braje, Wendy L; Kersten, Daniel; Tarr, Michael J; Troje, Nikolaus F: Illumination effects in face recognition. Psychobiology, 26(4):371–380, 1998.

[DER16] Dantcheva, Antitza; Elia, Petros; Ross, Arun: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Transactions on Information Forensics and Security, 11(3):441–467, 2016.

[EEH14] Eidinger, Eran; Enbar, Roee; Hassner, Tal: Age and Gender Estimation of Unfiltered Faces. IEEE Transactions on Information Forensics and Security, 9(12):2170–2179, 2014.

[Ha21] Hassan, Bilaland Izquierdo, Ebrouland Piatrik Tomas: Soft biometrics: a survey. Multimedia Tools and Applications, Mar 2021.

[LH15] Levi, Gil; Hassner, Tal: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 34–42, 2015.

[Ma18] Masood, Sarfaraz; Gupta, Shubham; Wajid, Abdul; Gupta, Suhani; Ahmed, Musheer: Prediction of human ethnicity from facial images using neural networks. In: Data Engineering and Intelligent Computing, pp. 217–226. Springer, 2018.

[Ni15] Nixon, Mark S; Correia, Paulo L; Nasrollahi, Kamal; Moeslund, Thomas B; Hadid, Abdenour; Tistarelli, Massimo: On soft biometrics. Pattern Recognition Letters, 68:218–230, 2015.

[Zh17] Zhang, Ke; Gao, Ce; Guo, Liru; Sun, Miao; Yuan, Xingfang; Han, Tony X; Zhao, Zhenbing; Li, Baogang: Age group and gender estimation in the wild with deep RoR architecture. IEEE Access, 5:22492–22503, 2017.