

Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race

Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, Bernard J. Jansen

Qatar Computing Research Institute, HBKU
HBKU Research Complex, Doha, P.O. Box 34110, Qatar
{sjung,jan,hkwak,jsalminen,bjansen}@hbku.edu.qa

Abstract

In this research, we evaluate four widely used face detection tools, which are Face++, IBM Bluemix Visual Recognition, AWS Rekognition, and Microsoft Azure Face API, using multiple datasets to determine their accuracy in inferring user attributes, including gender, race, and age. Results show that the tools are generally proficient at determining gender, with accuracy rates greater than 90%, except for IBM Bluemix. Concerning race, only one of the four tools provides this capability, Face++, with an accuracy rate of greater than 90%, although the evaluation was performed on a high-quality dataset. Inferring age appears to be a challenging problem, as all four tools performed poorly. The findings of our quantitative evaluation are helpful for future computational social science research using these tools, as their accuracy needs to be taken into account when applied to classifying individuals on social media and other contexts. Triangulation and manual verification are suggested for researchers employing these tools.

1 Introduction

Computational social science is increasingly employing tools for face detection from images. This capability extends the potential of social media analytics by, for example, inferring the gender and age, for example, from a user’s photograph or platform profile picture. The general purpose is to enable large-scale analyses of social media users employing inferred user demographics. Face detection tools are essential for large-scale inference of demographic attributes, since it is difficult to achieve this via other methods, such as manual labeling. Yet, there are key concerns about the underlying accuracy of these facial recognition tools. As a practical example, in 2015, Google apologized for their photo apps having tagged African Americans as gorillas¹. In 2016, the passport application of an Asian man was rejected because the software mistakenly claimed his eyes were closed².

In this sense, the facial detection tools that are widely used in computational social science research require comprehensive evaluation for reliability. So far, these tools have been

validated in an ad-hoc manner; face detection results from sample data are compared with crowdsourced labels within a single research paper, for example. As this verification is usually confined to the type of data and tool used in an individual study, it is a justification of that study rather than a comprehensive analysis of the accuracy of available face detection tools. In this research, we provide a comprehensive measurement of four widely-used face detection tools employing multiple datasets to gauge their accuracy. Our results offer guidelines for future studies using these tools in computational social science research.

2 Related Work

Inferring the demographics of social media users is a typical concern in online research projects, as this information is often not readily available (An and Weber 2016). A popular method is to leverage user profile images (Wang, Li, and Luo 2016) to infer race, gender, and age from facial features. Facial analysis is becoming easier to use through commercial services and APIs, such as the ones evaluated in this research. In particular, Face++ has been used in a variety of studies (An and Weber 2016). For example, Chakraborty et al. study the contribution of demographic groups to Twitter’s trending topics using Face++ (Chakraborty et al. 2017). An and Weber collect profile images of 350K Twitter users, using Face++ to infer gender, age, and race for studying hashtag use by different demographic groups (An and Weber 2016). Garcia et al. study the correlation between the gender of a video uploader and the number of interactions using the inferred gender of the uploader via Face++ (Garcia, Abisheva, and Schweitzer 2017). Jung et al. investigate inferring demographic attributes from social media profile pictures and find that automatic facial recognition is problematic, prompting further research in this area (Jung et al. 2017). Overall, our research provides a comprehensive analysis of state-of-the-art tools, which is useful for future studies employing facial recognition tools for inferring demographic attributes.

3 Widely-Used Face Recognition Tools

Here, we compare four widely-used face detection tools in terms of their accuracy: Face++³, IBM Bluemix Vi-

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/>

²<https://goo.gl/nng4jf>

³<https://www.faceplusplus.com/>

sual Recognition⁴, AWS Rekognition⁵, and Microsoft Azure Face API⁶. We exclude Google Cloud Vision API⁷ because it does not support inferring gender.

	Face Attributes	Free Quota & Pricing
Face++	age with range gender ethnicity	1 image / second \$0.0001 / image
IBM	age (maximum/minimum) gender	250 images / day \$0.004 / image
Amazon	age (high/low) gender	5k images / month \$0.001 / image
MS	age gender	30k images / month \$0.0015 / image

Table 1: Overview of four popular face detection tools using in this research.

Table 1 summarizes the capabilities and features of each tool. All tools infer the attributes of age and gender, and Face++ can detect ethnicity as well. However, its ethnicity classification is limited to white, black, or Asian; thus, it is actually attempting to detect race, not ethnicity, as stated. Within a race, there can be many ethnicities, e.g. “Asian” can include “Korean” and “Japanese.” Regarding age, Face++ returns an exact inference and a range; IBM and Amazon return a maximum / high age and minimum / low age without a precise age calculation. MS infers an exact age, with a decimal point, without a range. For Gender, IBM and Amazon return the gender of the face with a confidence score.

4 Benchmark Datasets

To evaluate the accuracy of the face detection tools, we prepare multiple datasets to cover a wide range of genders, ages, and races, as well as the quality of the photos. We define accuracy as the degree to which the results of the tools conform to the correct value. The three datasets used in this evaluation are shown in Table 2. Noisy data refers to images of low quality or where the frontal of the face is not fully revealed in the image due to, for example, a helmet being worn or only the side of a face being exposed. In the following subsections, we briefly explain each of the three datasets.

	Clean	Noisy
Ground-truth	(1) 100 Celebrities	(2) IMDb-Wiki [†]
No ground-truth	-	(3) Twitter profiles

[†]gender and age but no ethnicity information are available

Table 2: Typology of the datasets used in the evaluation of the four facial recognition tools.

⁴<https://www.ibm.com/watson/services/visual-recognition/>

⁵<https://aws.amazon.com/rekognition/>

⁶<https://azure.microsoft.com/en-us/services/cognitive-services/face/>

⁷<https://cloud.google.com/vision/>

4.1 100 Celebrities Dataset

We collect images of famous celebrities with the ground-truth demographic information from FamousBirthdays.com⁸. In addition to the photos, we also collected the birthday of the celebrity and the date when the photo was taken. The reason behind choosing photos of celebrities is due to the generally high quality of the images and the existence of ground truth of age, gender, and race. We carefully create a stratified random samples of 100 celebrities that consists of 33 whites, 33 blacks, 34 Asians, with balanced age groups of each: 11 young (13 to 34), 11 middle-aged (35 to 54), and 11 old (55 or older). In addition, we balance out gender groups as 47 female and 53 male. We double check the demographic information provided via FamousBirthdays.com with other external sources. As Table 3 shows, we have very clear images of celebrities along with their age, gender, and racial information.

			
20 female white	40 male white	63 female black	39 male Asian

Table 3: Examples from ‘100 celebrities’ with ground truth.

4.2 IMDb-Wiki Dataset

We employ the IMDb-Wiki dataset, which contains 52,489 photos introduced by a prior study (Rothe, Timofte, and Gool 2016). The dataset provides metadata containing the birth date, year when the photo was taken, gender, name, and so on. For each of the images, we use age and gender as ground truth. We determine age using the birthday and the year when the photo was taken. Table 4 presents a few examples from this dataset. While the dataset of 100 celebrities has clear and high-quality headshots, the images in this dataset are noisy (e.g., not of square size, in grey-scale, or having several faces, etc.). Thus, the images in this dataset are likely to be more challenging for face recognition tools to use.





			
1967/02/12 2009 male	1921/11/03 1966 male	1973/02/04 2006 male	1986/06/13 2011 female

Table 4: Examples from ‘IMDb-Wiki’ with ground truth.

⁸<https://www.famousbirthdays.com/>

4.3 Twitter Profile Photo Dataset

For this dataset, we use the profile images of actual Twitter users. Using Twitter REST APIs, we leverage retweets and replies associated with the tweets published by a major media corporation. We obtained 10,309 images from distinct user profiles and chose 1,000 randomly.

To establish the ground truth for these images, we use crowdsourcing. We assign three workers for each profile image and ask three questions. First, we ask whether there is a face, or faces, in a given photo. Second, we ask whether a face is a male or female. Third, if there is a face, we ask for the age. Considering the difficulty of determining the exact age, we offered a list of age bins, specifically (13-17), (18-24), (25-34), (35-44), (45-54), (54-64), and (65+). These age bins are aligned with those employed by most of the social media analytics tools (e.g., Facebook, Google Analytics).

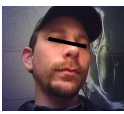



			
25-34 male	18-24 female	55-64 male	18-24 female

Table 5: Examples from Twitter dataset with ground truth established by crowdsourcing. Note: Black bars added.

From Twitter profiles, we collected 165 images with ground truth for age, and 592 images with ground truth for gender, where three workers reached an agreement for age and gender. The other images either did not contain a face, contained multiple faces, or there was no agreement on gender or age. Among these, we use 94 (called Twitter-Age) and 373 (called Twitter-Gender) images. Table 5 shows a sample of Twitter profile images with ground truth. Getting three agreements from CrowdFlower workers implies that the profile images are humanly recognizable and classifiable relative to the other profile images that did not reach an agreement. The small number of usable Twitter profile images illustrates the practical issues of using social network profile images for facial recognition.

5 Measurement Tasks

So as to evaluate the detection tools, our measurement consists of four tasks, which are:

- **Face Detection:** We evaluate whether a tool can correctly detect a face in a given photo, a basic task for a facial detection tool.
- **Gender Detection:** We evaluate whether a tool can correctly infer gender from a given face in an image. All of the tools support gender from detected faces, which all define a face as being either male or female.
- **Age Detection:** We evaluate whether a tool can correctly infer age from a given face. All of the tools provide age inference from detected faces, but each tool has a different scale for age, such as exact age (MS), age range (Face++), and minimum and maximum age (IBM and Amazon). For

a comparing accuracy, we compute the bin-level accuracy (e.g., 18-24, 25-34). Moreover, we compute the average minimum and the average maximum errors of the inference.

- **Race Detection:** We evaluate whether a tool can infer race from a given face. Although race detection is only supported by Face++ to indicate a face is white, black, or Asian, it is an increasingly investigated feature in computational social science research (An and Weber 2016).

6 Measuring the Accuracy

6.1 Face Detection Evaluation

We focus on images with only one face, which is a typical form of profile photos. The result of the face detection task summarized in Table 6. All the tools accurately detect faces when there is an image with clear headshot, such as those in the 100 celebrities and Twitter profile datasets. However, when images are noisy, such as in the IMDb-Wiki dataset, the performance of each tool drops to a high of 62.4% (Face++).

Dataset	Face++	IBM	Amazon	MS
100 celebrities (100 imgs)	100% (100)	98% (98)	99% (99)	99% (99)
IMDb-Wiki (52,478)	62.4% (32,769)	79.7% (41,811)	76.8% (40,296)	71.2% (37,343)
Twitter-Age (94)	100% (94)	91.5% (86)	92.6% (87)	90.4% (85)
Twitter-Gender (373)	100% (373)	93.6% (349)	91.7% (342)	90.1% (336)
All dataset (53,045)	62.8% (33,336)	79.8% (42,344)	76.9% (40,824)	71.3% (37,863)

Table 6: Share of images where a face is detected.

Table 7 shows the percentage of images correctly identified by all four tools. Compared to Table 6, the percentage is considerably smaller. For example, in Twitter-Age, although the lowest accuracy of each tool is of MS (90.4%), the percentage of the profile photos that are identified correctly by all the tools is only 79.8%. This shows that the errors of the tools do not overlap, implying they have different classification strengths and weaknesses. This finding highlights the importance of triangulation in facial recognition research (i.e., using more than one facial recognition tool).

6.2 Gender Detection Evaluation

We then compare the inferred gender from the detected faces with the ground truth. As presented in Table 8, all the tools, except IBM, have an accuracy of more than 0.9. Amazon and MS are good at inferring gender for images with clear headshots relative to the other tools. MS has the highest accuracy for gender among the four tools. We found that IBM returns a score of 0.0 for gender for several images in all datasets, even though it detects the face and determines age. In these cases, IBM always classifies the instance as woman.

Dataset	Percentage
100 celebrities (100 imgs)	97% (97)
IMDb-Wiki (52,478)	52.7% (27,641)
Twitter-Age (94)	79.8% (75)
Twitter-Gender (373)	80.4% (300)
All dataset (53,045)	52.9% (28,113)

Table 7: Share of images where face detected for all tools.

We counted these as incorrect since crowdsourcing reached an agreement to infer their genders.

Dataset	Face++	IBM	Amazon	MS
100 celebrities	1.00	0.95	0.99	1.00
IMDb-Wiki	0.92	0.7	0.95	0.97
Twitter-Gender	0.93	0.78	0.98	0.99
All dataset	0.92	0.7	0.95	0.97

Table 8: Results of gender inference evaluation.

6.3 Age Detection Evaluation

When it comes to inferring to age, all tools have a difficulty correctly classifying. From Table 9, we see that MS has the highest overall accuracy in detecting age, but the accuracy still stays at less than 0.5.

Dataset	Face++	IBM	Amazon	MS
100 celebrities	0.28	0.53	0.35	0.37
IMDb-Wiki	0.34	0.37	0.3	0.45
Twitter-Age	0.41	0.36	0.36	0.66
All dataset	0.34	0.37	0.3	0.45

Table 9: Results of age inference evaluation.

Table 10 presents [the average minimum errors, the average maximum errors] of age inference by each tool. The error is defined as [the inferred age] - [the ground truth]. If the error is positive, age is overestimated. If negative, age is underestimated. As MS does not provide an age range, MS has the same values for both. Since Twitter-Age dataset does not have the ground truth for the age, we excluded it. We found a large error range for age inference. In the 100 celebrities dataset, on average, celebrities are underestimated by 15.2 years. Amazon also shows a major error range. Compared to other tools, MS shows a stable range of errors in age; however, all tools tend to overestimate the young age group.

6.4 Race Detection Evaluation

Here, we use the 100 celebrities dataset as it is the only dataset with ground truth of race. The accuracy of inferring race by Face++ is 0.93. This high accuracy might be the result of the high-quality image dataset.

Dataset	Face++	IBM	Amazon	MS
100 celebrities	[-15.2,-1.6]	[-8.9,-1.3]	[-10.9,6.9]	[-5.2,-5.2]
IMDb-Wiki	[-8.9,6.4]	[-5.5,1.6]	[-2.1,16.6]	[2.7,2.7]
All dataset	[-8.9,6.4]	[-5.5,1.6]	[-2.1,16.5]	[2.7,2.7]

Table 10: Average errors of age inference.

7 Discussion and Implications

The results of our accuracy evaluation using multiple datasets for four popular facial recognition tools highlight the need for triangulation as a crucial step for better computational social science research, even for the relatively simple task of face detection within an image. Reviewing our specific results we see a trend of high accuracy for gender, with three of the tools performing with an accuracy of above 90 percent for all datasets. Concerning race, only one tool offers this capability, Face++, and the accuracy is quite high, above 90 percent. However, this was evaluated by using a high-quality dataset of images. Future research is needed to determine if such accuracy holds for noisy images.

All of the tools performed poorly for age, even with the relaxed task of determining an age bin instead of exact age. Moreover, the average error regarding age for all tools was quite high. We conjecture that an individual’s age may be a difficult attribute for facial recognition tools to discern, perhaps due to cosmetic surgeries, the use of make-up, hair coloring, etc. As observed from crowdsourcing, age is difficult for even humans to discern. In future research, we are considering a more nuanced evaluation of these tools, including larger datasets and investigation into subgroups of gender, age, and race.

References

- An, J., and Weber, I. 2016. #greysanatomy vs. #yankees: Demographics and hashtag use on Twitter. In *ICWSM*, 523–526.
- Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *ICWSM*.
- Garcia, D.; Abisheva, A.; and Schweitzer, F. 2017. Evaluative patterns and incentives in YouTube. In *International Conference on Social Informatics*, 301–315. Springer.
- Jung, S.-G.; An, J.; Kwak, H.; Salminen, J.; and Jansen, B. J. 2017. Inferring social media users demographics from profile pictures: A face++ analysis on twitter users. In *Proceedings of 17th International Conference on Electronic Business*.
- Rothe, R.; Timofte, R.; and Gool, L. V. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*.
- Wang, Y.; Li, Y.; and Luo, J. 2016. Deciphering the 2016 US presidential campaign in the Twitter sphere: A comparison of the Trumpists and Clintonists. In *ICWSM*, 723–726.