BUILDING A FACE DATABASE OF ARAB FACES TOWARD EVALUATING BIAS IN FACIAL ANALYSIS SYSTEMS

Ashraf Khalil[1], Suha Glal[1], Khider Ahmed[2], Sana Zeb Khan[3] and Aysha Abdulgani[4]
[1]*College of Technological Innovation, Zayed University*
Abu Dhabi, UAE
[2]*Mechatronics Engineering Graduate Program, American university of Sharjah*
Sharjah, UAE
[3]*Department of Computer Science and Engineering, American university of Sharjah*
Sharjah, UAE
[4]*United Arab Emirates University*
Al Ain, UAE

ABSTRACT

Machine learning algorithms are fundamentally driven by the data provided by humans; consequently, the decisions made by those algorithms are not free from human bias. This is particularly evident in the case of facial analysis systems that employ machine learning algorithms. Recent studies have shown that the decisions made by many of the commercially available facial analysis systems are prejudiced against certain groups of race, ethnicity, age, gender and culture. Further studies have identified that the underlying reason for such biased decisions is that the open source material available for facial image databases which are used in commerce and academia to train the algorithms has meager diversity in these categories. To compound this issue, facial analysis technology is promoted by influential companies and artificial intelligence service providers without affirming the fairness and accuracy of the decisions given by these systems. To minimize bias and ensure representation of the Middle Eastern population in the imminent growth of this technology, we propose the development of two Arab face databases along with an algorithmic audit involving seven commercially available facial analysis systems. Of the databases, the first, Arab-LEANA, will include 300 Arab subjects' face images with variation in lighting, expression, accessory, nationality and age (LEANA). The second, Arab Public Figures Faces (APFF), will contain images and videos of 300 Arab public figures captured "in the wild". Faces for APFF will be selected manually from the internet since manual selection of faces will result in a high degree of variability in scale, pose, expression, illumination, age, occlusion and make-up. These databases will provide the worldwide community of face recognition researchers with a large-scale, diverse collection of Arab face images for training and evaluating algorithms toward developing a more representative, and therefore more robust, capacity for facial analysis. This, in turn, will facilitate the development of more accurate face recognition technology as it prepares to go mainstream and enter numerous facets of modern life.

KEYWORDS

Arab Face Database, Arab Public Figures Face, Algorithmic Audit, Image Manipulation, Image Processing, Machine Learning

1. INTRODUCTION

Open-source image databases are used for building, training and testing facial analysis software that are used by researchers in academia (Escalera et al, 2016) (Huang et al, 2008) (Klare et al, 2015). However, these open-source databases usually lack diversity, and building a database is an expensive and time-consuming process. Databases such as MORPH (MORPH, n.d.) are built from source materials that lack diversity in terms of race, culture and ethnicity. Escalera et al. have discussed the issues connected with limitations of databases that have a misrepresentative range of facial photos (Escalera et al, 2016). These databases, despite their high degree of skew, are largely utilized for building face recognition software. For example, the dataset IJB-A (Klare et al, 2015) with 500 subjects, containing more than 5712 entries, was developed for unconstrained face

detection and recognition. The subjects used in the work had a demographic distribution of Europe (149), North America (135), Asia (89), South America (50), Middle East (29) Oceania (7), and Africa (4). Similarly, the MS- Celeb-1M (Guo et al, 2016) database that was built for the same purpose as IJB-A has 100,000 subjects with 10 million images. The top 5 represented countries of the subjects were the United States, Great Britain, Germany, Canada and France. Similar databases such as LFW (Huang et al, 2008), PubFig (Kumar et al, 2009), Faces (Ebner et al, 2010), CMU-Pittsburgh AU-Coded Face Expression (Kanade et al, 2000) clearly show that the Middle East and North African region (Arabic-speaking people) are underrepresented.

On top of that, despite being the world's second largest religious group, Muslims are generally underrepresented in face databases. Face accessories are very common among Muslims. It is well known that many female Muslim women wear a head cover and many wear a face cover as well. In situations such as face scans at airport departure gates, such women may be required to temporarily uncover their face in the interest of security. However, in scenarios where video surveillance is being used at shopping malls for security or for target marketing, automatic facial analysis will perform poorly in the case of many Muslim women. Figure 1 illustrates some of the very commonly used face accessories.

We predicate that these accessories will be challenging to current state-of-the-art automatic facial analysis or face recognition systems. In this work, we propose the development of two Arab face databases along with an algorithmic audit involving seven commercially available facial analysis systems. The first database, Arab–LEANA, will contain 300 facial images of Arab subjects with variations in lighting, expressions, accessories, nationality, and age. The second database, APFF, will consist of 300 images capturing Arab public figures “in the wild”. To the best of our knowledge, it is the first time ever to build a facial database of Arab subjects.

Figure 1. Face cover (called niqab, burqa, or gashowa) worn in various Arab and Muslim countries. (a) represents face cover worn mostly in UAE, KSA, Yemen, and other Gulf countries, (b) Jordon, (c) Afghanistan, (d) Egypt. (e) Traditional head cover for men in some Arab countries

In recent times, researchers working on facial analysis systems have created databases to represent faces of a particular ethnicity or category such as the Indian Face Database (Vidit and Amitabha, 2002), Indian Movie Face Database (Setty et al, 2013), Japanese Female Facial Expression (JAFFE) Database (Lyons, 1998), CAS-PEAL Chinese Face Database (Gao et al, 2007) and CaNAFF Male North African faces (Courset et al, 2018). This is a welcoming move and we applaud the research teams as they play a pivotal role in advancing the accuracy of facial analysis technology. These research groups are considering factors such as race and ethnicity.

The first aim of our project is to create two Arab face benchmark databases. The first database, Arab LEANA, will include Arab subjects' face images (N=300) with different sources of variation, especially lighting, expression, accessories, nationality and age (LEANA). The second database, Arab Public Figures

Faces (APFF), contains images and videos for 300 Arabic public figures taken “in the wild” (without posing). After labeling, all the subjects will be manually localized with bounding boxes for face detection, as well as fiducial nodal points for the center of the two eyes and base of the nose (if visible).

The second goal is to perform an algorithmic audit using the two developed Arab face benchmarks. The audit will evaluate the performance of seven commercial face analysis systems and report the results. Specifically, the audit will include Microsoft, IBM, Face++, Amazon Rekognition, DeepFace, Kairos and Facelytics face recognition software. These softwares were chosen because these companies have heavily invested in artificial intelligence, secured substantial market shares in the machine learning services domain, and provide public demonstrations of their facial analysis technology. Also, six of the seven chosen software have already been involved in a similar algorithmic audit (Raji and Buolamwini, 2019). Therefore, it would be possible for us to compare or at least reference their previous performance in our study. This will be the first algorithmic audit that will examine seven commercial face recognition systems performance in reference to cultural differences.

2. METHODOLOGY

In order to mitigate bias against Arab subjects (both male and female) of Middle-Eastern and North-African regions, we need a large set of unconstrained Arab face images. In this section, the methodology that will be adopted for building an Arab facial image database is presented. The procedure for collection and annotation of Arab Public Figures Faces (APFF) database was motivated by the work of Klare et al. (2015).

2.1 Arab Face Benchmark Databases

2.1.1 Development of Arab-LEANA Database

The first database - Arab LEANA (Lighting, Expression, Accessory, Nationality and Age), will consist of 300 Arab subjects’ face images (150 females and 150 males). These images will be collected by taking pictures of Arab subjects with variation in Lighting, Expression, Accessory, Nationality and Age. A brief description of the variations is given below.

- Lighting: Variability in lighting is achieved by positioning 9 lamps in different directions while capturing the image.
- Expression: Each subject gives 6 facial expressions- neutral, happiness, sadness, disgust, fear, and anger.
- Age: Images include a range from young, middle-aged and elderly age groups.
- Nationality: Subjects are from the Middle-Eastern and North-African region - Algeria, Comoros, Djibouti, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Mauritania, Oman, Palestine, Qatar, KSA, Somalia, Sudan, Syria, Tunisia, UAE, Yemen.
- Accessories: 6 kinds of accessories commonly in use among Arab men and women are used. This includes (i) head-scarf, (ii) Niqab – face veil, (iii) Agal – head band, (iv) Ghutra – Muslim male attire (v) Burqa – Muslim female attire (vi) Shemagh or Keffiyeh – scarf popularly used by Arabs.

To shoot the Arab subjects’ faces, we will set up a special photographic studio room in our computer science laboratory at Zayed University, UAE and the necessary devices will be arranged in the studio: precisely, a camera system, a lighting arrangement, accessories, and various backgrounds. By switching on and off each lamp while the room lights are kept on, altered directional lighting conditions are simulated. A switch matrix will be utilized to control the on/off settings of these lamps.

The sketch map of the cameras’ distribution on the semicircle arm and the lamps’ location are shown in Figure 2. To generate different lighting conditions, a lighting system will be set using nine fluorescent lamps as shown in Figure 2 (a). As shown in Figure 2 (b), for each subject, seven cameras will be spaced equally in a horizontal semi-circular shelf setup to simultaneously capture images across different poses in one shot. Also, each subject will be asked to look up and down to capture 14 images in another two shots. The camera system will consist of seven digital cameras connected to a computer via USB interface specially configured to support them. We will develop software to control the seven cameras and capture the images in one shot. Images will

be stored in a hard drive using a uniform naming convention. Additionally, a professional digital media design student from ZU will be recruited to post-process and edit the images.

In a subsequent validation study, each face will be rated in terms of facial expression and perceived age by young, middle-aged, and elderly females and males (N=300, 50% male). The rating serves two purposes: First, to validate the database and, second, to provide reference for researchers (ground truth). In other words, the aggregated rating will serve as a degree of certainty or confidence interval that the emotion or age presented in the image is perceived in the same way by the raters.

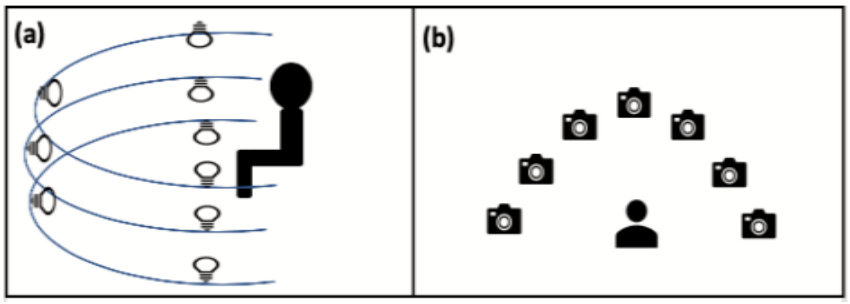

Figure 2. Setup in the photographic room with devices: (a) Lamps positions and configuration (b) Cameras positions

Given the variation we have in the image shooting procedure (ex. pose, expression, age and lighting), we believe 300 subjects for the validation study will be necessary. Also, studies have reported a high turnover among raters due to the difficulty of the task therefore we would like to account for the turnover rate by recruiting a large sample initially. A simple software will be developed for the pictures rating task. Participants of the photo shoot will be compensated with 300 AED each while participants of the validation study will also be compensated with 300 AED each.

2.1.2 Development of Arab Public Figures Face (APFF) database

The second database, Arab Public Figure Faces (APFF) will contain images and videos of 300 Arab public figures captured “in the wild”. Faces for APFF will be selected manually from the internet since manual selection of faces will result in a high degree of variability in scale, pose, expression, illumination, age, occlusion and make-up.

The database will be built through the following steps:

1. Subject Specification: Subjects for the database will be manually specified, ensuring a diverse representation across Arab countries.
2. Image and Video Retrieval: Images and videos of the subjects will be located by searching the internet for Creative Commons licensed content. The subject's name, URL, and relevant details will be stored in an Excel file.
3. Scraping and Storage: A scraping software will automatically download the subject's images and save all information in a relational database. For videos, clipped versions and extracted frames will be saved.
4. Annotation: Annotation will be performed using Amazon Mechanical Turk (AMT), a crowdsourcing service. This process will involve creating precise ground truth data, including bounding box positions for faces, subject labels for face recognition, and facial nodal points for research purposes. Additional metadata such as skin tone, sex, facial pose, occlusion, and ambience will also be collected. The annotation task will be executed as given below:

- Errand 1: Annotate bounding boxes around faces in images and video frames, ensuring each head is accurately enclosed. Multiple people in an image will have individual bounding boxes. Five AMT workers will annotate each image, and their annotations will be consolidated into a single set.
- Errand 2: Determine the bounding box corresponding to the person of interest (POI) after annotating all faces. This step helps identify the POI in images with multiple persons. Annotators will compare a reference image with the target image and select the appropriate bounding box. Additional face localization information, such as eye centers and nose base, will be annotated.
- Errand 3: An expert analyst will thoroughly review and manually correct any annotation errors in the consolidated annotations.

Considering variations in image capture (pose, expression, age, lighting), 300 subjects are needed for the validation study. To account for high turnover among raters, a large initial sample will be recruited. A user-friendly software will be developed for picture rating, and the database will be publicly accessible.

2.2 Algorithmic Audit Using Arab Face Benchmark Databases

Researchers who have worked in developing a database of a specific race use their database to evaluate the performance accuracy of state-of-the-art facial analysis software that are commercially available (Courset et al, 2018) (Ebner et al, 2010). In our work, too, we will evaluate the performance accuracy of seven commercial gender classifiers, namely Microsoft Face (Microsoft Azure, n.d.), IBM (IBM Diversity in Faces, n.d.), Face++ (Megvii, n.d.), Amazon Rekognition (Amazon, n.d.), DeepFace (Deepface, n.d.), Kairos (Kairos, n.d.) and Facelytics (Facelytics, n.d.) on our APFF database and Arab LEANA face database. For the auditing purpose, the overall classification accuracy, male classification accuracy, and female classification accuracy will be measured and the positive predictive value (PPV) will be computed. We then evaluate the true positive rate, false positive rate, and error rate (1-PPV) of the following groups: all subjects, male subjects, female subjects, males with traditional accessories, and females with traditional accessories. The performance metrics will reveal whether the databases used in developing these popular facial analysis systems are biased or not.

3. CURRENT STATUS AND FUTURE WORK

The Arab Face Database (AFD) comes with detailed annotations in terms of age, country of origin, and gender. As a first step towards developing AFD, we have already started building the APFF database. The procedure of building APFF database is discussed below. To start with, the famous Arab subjects that are to be included in the database were specified manually. While giving specifications, we took care to ensure that the country of origin of the subjects is well distributed across the Arab region. We have manually chosen Arab public figures from 22 different Arab countries. The subjects are famous in various fields such as business, literature, media, sports, entertainment, politics, religion, & philanthropy. The subjects chosen include a wide range of professionals such as actor, comedian, athlete, music composer, singer, diplomat, director and ceo of companies, judge, jurist, member of royal family, researcher, scholar, scientist, social workers, philanthropist, vocalist, writer, poet, emir, fashion designer, economist, news presenter, model, political party leader, former and present president, prime minister, journalist, minister and military commander. A total of 604 Arab subjects were selected of which 277 are females and 327 are males. A reference image database was manually built by selecting one good quality image of each of the 604 subjects. This database will be given to annotators as a reference image for each subject. Table 1 shows the diversity of the Arab subjects in the built database in terms of country of origin, gender and the various fields in which they are famous.

Table 1. Demographics of Arab subjects in the APFF database in terms of country of origin, gender, professional fields

		Gender		Field					
Country of Origin	No. of subjects	Female	Male	Philanthropy	Business	Litera-ture	Media, Sports & Entertainment	Politics	Religion
Algeria	30	15	15	0	0	4	20	5	1
Bahrain	30	15	15	0	1	4	19	6	0
Comoros	15	4	11	0	0	0	11	4	0
Djibouti	16	0	16	0	0	1	6	9	0
Egypt	30	15	15	0	1	8	14	5	2
Iraq	30	15	15	0	0	1	21	6	2
Jordan	30	15	15	2	2	4	15	7	0
Kuwait	30	15	15	0	2	1	19	6	2
Lebanon	30	15	15	2	0	2	20	6	0
Libya	23	8	15	3	0	1	8	11	0
Mauritania	24	9	15	3	1	2	3	13	2

Morocco	30	15	15	0	0	4	20	5	1
Oman	23	8	15	0	4	1	12	5	1
Palestine	30	15	15	0	1	4	19	6	0
Qatar	30	15	15	0	1	8	14	5	2
KSA	30	15	15	0	0	1	21	6	2
Somalia	30	15	15	2	2	4	15	7	0
Sudan	30	15	15	0	2	1	19	6	2
Syria	30	15	15	2	0	2	20	6	0
Tunisia	23	8	15	3	0	1	8	11	0
UAE	30	15	15	2	2	6	14	5	1
Yemen	30	15	15	5	0	6	8	11	0
All subjects	**604**	**277**	**327**	**24**	**19**	**66**	**326**	**151**	**18**

After manually compiling the list of Arab public figures' names, we located their images. The images and videos of the subjects were located by searching Google images with Creative Commons license. We used a web scraping tool called Octoparse that automates the process of locating the images. The tool takes in the list of subjects' names and, for each subject, it saves the URLs in a database under the subject's name. As the search is for Arab subjects, Google search in Arabic language resulted in more Creative Common (CC) images compared to the search conducted in English. The time taken by the software to scrape the data varied from subject to subject. This is because the search result for each subject is different. On the whole, it took approximately 8 hours to complete the scraping for search results in English and 13 hours for search results in Arabic.

The minimum number of URLs collected for a subject is zero as some of these URLs contain only the mention of the subject and do not have the subject's image. Certain subjects' images appear in a greater number of URLs, for instance, a subject's image appeared in 27 URLs. This implies that we may have as many as 27 images for a subject and on the other hand, we may not have any image for a subject in the database. Hence cleaning the database, that is removing those URLs that do not contain the subjects' images, has to be done and this has to be done manually.

We performed bulk downloading of the images from the set of URLs and saved them under each subject's name. The downloaded image files are saved under the subject's name in a sequential numerical order, such as Abdelmadjid Tebboune_1, Abdelmadjid Tebboune_2. This way, we have built a database of approximately 11,000 images of the 604 subjects, with nearly 6000 images resulting from search in Arabic and 5000 images from search in English.

In case of videos, a clipped version from the original codec and the extracted I-frames, have to be stored. In order to get the images from YouTube videos, we are working on finding a YouTube video scraping tool that can automate the process of collecting various frames/clips of the subject.

After generating the database of images, the next step is to perform annotation of the images. This will be performed through the crowdsourcing service given by Amazon Mechanical Turk (AMT) workers. We will give the annotation task to five AMT workers. We are preparing detailed visual and written guidelines on annotation for the AMT workers. As a part of the guidelines, the annotators will be provided with a good quality reference image of each of the subjects. If an image or video frame contains multiple persons, this reference image will help the annotators to locate the person of interest. Open-source image annotation software available for academic projects and commercial applications can be used to annotate the images. We found an open-source annotator, VGG Image Annotator (VIA), to annotate the images. We will use this VIA to annotate a sample image and the annotated image will be included in the visual guidance. In the guidelines, the AMT workers will be informed to use the same annotator to annotate our image database. The final step after annotation is consolidation of the annotations of the five AMT workers. We are working on finding a code to consolidate the annotations obtained from all the five AMT workers to obtain a label for each subject's image.

The second database, Arab–LEANA will be built using facial images of Arab subjects taken in a purposely-designed photoshoot. For this, the photo studio with arrangements shown in Figure 2 will be set up in the computer science laboratory at Zayed University.

After developing the two databases, we will perform the algorithmic auditing of seven well-known commercially available facial analysis softwares using our facial image database. We conducted a pilot study to evaluate the performance accuracy of Facelytics software in recognizing faces of Arab subjects wearing

traditional Arab face accessories. The results showed the vulnerability of the system in identifying the faces. The pilot study result is shown in Figure 3. This study will be extended to the other 6 facial analysis systems mentioned.

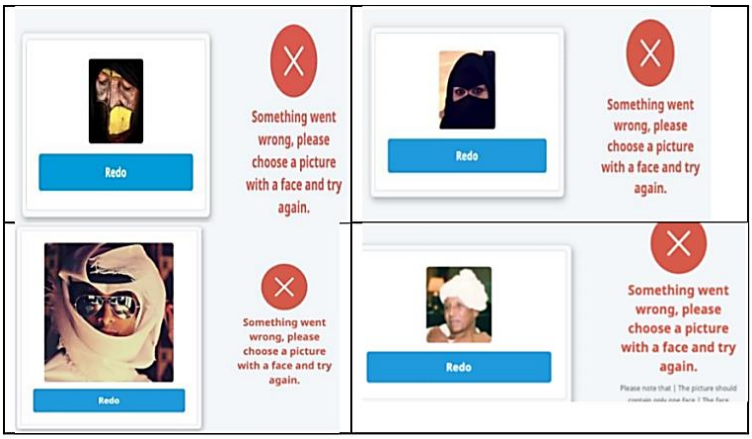

Figure 3. Results of pilot study conducted on Facelytic gender classification for 4 images of Arab Subjects.

4. CONCLUSION

We propose building two Arab facial image databases and conducting an algorithm audit on seven commercial facial recognition algorithms. The first database, Arab–LEANA, will consist of 300 Arab subjects' facial images with variations in lighting, expressions, accessory, nationality, and age. The second database, Arab Public Figure Faces (APFF), will include images of 604 Arab public figures captured from various sources.For the APFF database, we have already selected 604 Arab subjects, including 277 females and 327 males, and located their images using the web scraping tool Octoparse. The database currently contains approximately 11,000 images. We are in the process of selecting a tool to scrape videos from YouTube to gather additional clips or frames. The annotation of the images will be performed by five Amazon Mechanical Turk (AMT) annotators using clear guidelines and a reference image database prepared manually. The annotations will be consolidated to obtain a label for each subject. After completing the APFF database, we will proceed with building the Arab-LEANA database. These databases will enable an algorithmic audit of facial analysis software systems. The aim is to provide the research community with a diverse collection of Arab facial images for training and evaluating algorithms, ultimately enhancing the accuracy and robustness of facial recognition technology.

ACKNOWLEDGEMENT

This work was supported by the Abu Dhabi Award for Research Excellence (AARE) 2019 under Grant 274.

REFERENCES

Amazon, Amazon Web Services. Amazon Rekognition - Video and Image - AWS. Available at: https://aws.amazon.com/rekognition/ Accessed on 7th June 2021.

Courset, R. et al, 2018. The caucasian and north african french faces (CaNAFF): A face database. *International Review of Social Psychology*, *31*(1).

DeepFace: https://pypi.org/project/deepface/ Accessed on 7th June 2021.

Ebner, N.C. et al, 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, *42*, pp.351-362.

Escalera, S. et al, 2016. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-8.

Facelytics: https://www.facelytics.io/#try-it. Accessed on 7th June 2021.

Gao, W. et al, 2007. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(1), pp.149-161.

Guo, Y. et al, 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 87-102. Springer International Publishing.

Huang, G.B. et al, 2008, October. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

IBM Diversity in Faces. Available at: https://www.research.ibm.com/artificial-intelli%C2%ADgence/trusted-ai/diversity-in-faces/.

Kairos.: https://www.kairos.com/kairos-2.0/demos. Accessed on 7th June 2021.

Kanade, T. et al, 2000, March. Comprehensive database for facial expression analysis. In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, pp. 46-53. IEEE.

Klare, B.F. et al, 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1931-1939.

Kumar, N., Berg, A.C., Belhumeur, P.N. and Nayar, S.K., 2009, September. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*. pp. 365-372. IEEE.

Lyons, M.J., 1998. Facial expression database: Japanese female facial expression (jaffe) database. *http://www.kasrl.org/jaffe.html*.

Megvii, Face++ Facial Recognition Database. Available at: https://www.faceplusplus.com/face-detection/ Accessed on 7th June 2021.

Microsoft Azure, *Facial recognition.* Available at: https://azure.microsoft.com/en-in/products/cognitive-services/face/ (Accessed: 22 June 2023).

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.

Raji, I.D. and Buolamwini, J., 2019, January. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* , pp. 429-435.

Setty, S. et al, 2013, December. Indian movie face database: a benchmark for face recognition under wide variations. In *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)* (pp. 1-5). IEEE.

Vidit, J. and Amitabha, M., 2002. The Indian face database.