scientific reports

Check for updates

OPEN

Collection of 2429 constrained headshots of 277 volunteers for deep learning

Saki Aoto[1], Mayumi Hangai[2], Hitomi Ueno-Yokohata[3], Aki Ueda[4], Maki Igarashi[4], Yoshikazu Ito[5], Motoko Tsukamoto[6], Tomoko Jinno[4], Mika Sakamoto[1], Yuka Okazaki[7], Fuyuki Hasegawa[8], Hiroko Ogata-Kawata[9], Saki Namura[5], Kazuaki Kojima[9], Masao Kikuya[10], Keiko Matsubara[4], Kosuke Taniguchi[9] & Kohji Okamura[11]✉

Deep learning has rapidly been filtrating many aspects of human lives. In particular, image recognition by convolutional neural networks has inspired numerous studies in this area. Hardware and software technologies as well as large quantities of data have contributed to the drastic development of the field. However, the application of deep learning is often hindered by the need for big data and the laborious manual annotation thereof. To experience deep learning using the data compiled by us, we collected 2429 constrained headshot images of 277 volunteers. The collection of face photographs is challenging in terms of protecting personal information; we therefore established an online procedure in which both the informed consent and image data could be obtained. We did not collect personal information, but issued agreement numbers to deal with withdrawal requests. Gender and smile labels were manually and subjectively annotated only from the appearances, and final labels were determined by majority among our team members. Rotated, trimmed, resolution-reduced, decolorized, and matrix-formed data were allowed to be publicly released. Moreover, simplified feature vectors for data sciences were released. We performed gender and smile recognition by building convolutional neural networks based on the Inception V3 model with pre-trained ImageNet data to demonstrate the usefulness of our dataset.

Deep learning is often viewed as having been catapulted into fame by the significant achievements of AlexNet at the ImageNet Large Scale Visual Recognition Challenge in 2012[1]. AlexNet is a deep convolutional neural network (CNN) consisting of convolutional, max-pooling, fully connected, and softmax layers[2]. Long before this well-known breakthrough, a CNN with a back-propagation algorithm, namely LeNet, was proposed and implemented for the recognition of handwritten digits[3]. With the advancements in computer hardware performance, CNNs have become popular in computer vision as one of the most successful artificial intelligence methodologies[4].

In addition to the progress made in hardware and software, the contribution of big data to the success of deep learning has been substantial. Deep learning is based on a supervised machine learning algorithm that relies on a large quantity of labeled data. More than one million labeled photographs from ImageNet were used for AlexNet in 2012[2]. ImageNet is a dataset consisting of quality-controlled photos organized into more than 20,000 categories in a hierarchical manner[5]. It covers up to 15 million images, each of which has been hand-annotated. Without such collection and annotation, deep learning would not have progressed to the current level.

In the early days of this field, face recognition and gender classification tools were dependent on soft biometric traits that could be perceived by humans. SEXNET, which was the first neural network for gender classification,

[1]Medical Genome Center, National Center for Child Health and Development, Tokyo, Japan. [2]Department of Social Medicine, National Center for Child Health and Development, Tokyo, Japan. [3]Department of Pediatric Hematology and Oncology Research, National Center for Child Health and Development, Tokyo, Japan. [4]Department of Molecular Endocrinology, National Center for Child Health and Development, Tokyo, Japan. [5]Center for Regenerative Medicine, National Center for Child Health and Development, Tokyo, Japan. [6]Department of Genome Medicine, National Center for Child Health and Development, Tokyo, Japan. [7]Center for Maternal-Fetal, Neonatal and Reproductive Medicine, National Center for Child Health and Development, Tokyo, Japan. [8]BioBank, National Center for Child Health and Development, Tokyo, Japan. [9]Department of Maternal-Fetal Biology, National Center for Child Health and Development, Tokyo, Japan. [10]Department of Information Technology and Management, National Center for Child Health and Development, Tokyo, Japan. [11]Department of Systems BioMedicine, National Center for Child Health and Development, Tokyo 157-8535, Japan. ✉email: okamura-k@ncchd.go.jp

achieved accuracy of over 90%[6] although it learned only 90 facial images at that time. Since then, numerous facial datasets, such as FERET[7], JAFFE[8], Eigenflow Based Face Authentication (EBFA)[9], LFW[10], Adience[11], IMDB-WIKI[12], AFAD[13], UTKFace[14], AIST Face2017[15], and Flickr-Faces-HQ (FFHQ)[16], have been developed and released. The majority of these datasets imposed challenging tasks when using images of an unconstrained nature. Because still face images can readily be obtained from the Internet, IMDB-WIKI, for example, collected more than 524,000 images along with their biological age information. However, the copyright usually belongs to the original owners, and not to the research group. Formally collecting such headshots with agreement from each subject is a formidable task involving the management of personal information. In contrast to unconstrained images, constrained images may be analyzed to understand the unrecognized soft biometric traits of human faces.

In the current study, we first formed a study group for the data sciences, traversing many departments and backgrounds in our institute, the National Center for Child Health and Development (NCCHD). We endeavored to construct an online collection system of headshots and to release a dataset of annotated headshots for use in deep learning. The dataset includes not only two-dimensional images but also one-dimensional feature vectors that are beneficial for other data sciences. After agreement was obtained from the volunteers, the system requested them to upload image data without personal information. When photographs were captured, the pose and action were constrained. Deviated headshots were manually eliminated, or were rotated, scaled, and set in fixed positions. After reviewing each face, we voted on the gender and smiling labels, which were then determined by majority. By employing leave-one-out cross-validation on all of the released data, our CNN could achieve a volunteer-based accuracy of 98.2% in recognizing the gender of a person.

Results

Application to ethics review committee. Our initial members intended to perform deep learning by using our collected data. The idea was to surmise the gender from a headshot image or face photograph. We were interested in determining how accurately common deep learning methods could recognize the gender without individual bearings; that is, by using constrained photographs. When we consulted on this matter with the director of our research institute on July 5, 2018, he encouraged us to proceed and advised us to begin with applying for ethics review. The idea was approved with a number of 2045 on January 10, 2019. However, there were several restrictions. We decided not to collect the volunteer names, gender, and contact information including e-mail addresses. We provided the volunteers with a withdrawal right for 50 days after uploading their photographs. We were allowed to publicly release the data after trimming, size and resolution reduction, decolorizing, and converting them into array forms consisting of only pixel intensity data.

Preparation for data collection. The initial team constructed a common gateway interface (CGI) submission system using HTML5 and Python to obtain both informed consent and headshots from the volunteers. Our intention and purpose appeared first on the website, following which we clarified what we would and would not collect. Through this process, we also collected the submission timestamps, IP addresses, and original file names from the submitters for the purpose of identifying the volunteers. Volunteers were requested to straighten up, look straight ahead, take off glasses, hats, scarfs, mufflers, masks, and so forth, and use plain background for the photographs as much as possible. Make-up including wearing earrings was permissible. They were requested not to cover eyes with hair and not to cover their faces. The first inquiry confirmed whether he or she was ready to take his or her 9 photographs or already had 9 photographs captured for this purpose. We did not accept photographs taken for other purposes. The second confirmed whether he or she was the person concerned. The third confirmed whether he or she understood our motive and purpose, and agreed to submit photographs of their own. The fourth confirmed the conditions for the withdrawal of the agreement. Thereafter, the year of birth, with neither the month nor day, was required to confirm that the volunteer was no less than 20 years old. We did not collect data from any individual younger than 20 years of age.

Collection of volunteers and their headshot images. We first sent e-mails to all employees working at our institute, the NCCHD, on February 7, 2019. The first volunteer submitted his or her data the following day. Thereafter, we hosted several explanatory seminars and announced the recruiting of not only volunteers who could generously provide their headshot images, but also team members who wished take part in the study of data sciences. By the end of January 2020, our web server had received 282 submissions. We requested 9 face images per volunteer and obtained all images from 279 people, although we failed to obtain 1 and 9 images from one and two volunteers, respectively. All of the successfully uploaded image files were in JPEG format. The smallest image file was 54.2 KiB with 960×721 pixels, whereas the largest one was 5.03 MiB with 4160×3120 pixels. We used the Exif Orientation information and did not look at other metadata. The 9 uploaded files were investigated using a POSIX command **cmp** repeatedly to determine whether identical files existed. The redundancy of these files was eliminated by removing either of the identical files. Several submitters uploaded identical files twice or more in the 9-file set. Some complained to us that 9 image files were too many to handle using a smartphone. The calling for volunteers was not restricted to employees or to Japanese. Nevertheless, the majority of the volunteers appeared to be Japanese or East Asian. Although individuals who appeared South Asian or European were included, our collection was biased in terms of skin-type or racial balance[17]. Instead of collecting personal information, we instantly issued an agreement number to each volunteer to deal with subsequent withdrawal requests. However, we did not receive any such requests or inquiries.

Quality control of uploaded image files. When the team members flicked through the images, we noticed that one volunteer uploaded his or her data thrice and another uploaded his or her data twice. The second and third redundant submissions were deleted. We first employed OpenCV[18] face recognition to prepare

Figure 1. Examples of headshot data. The images were saved as 359 × 359 pixel squares with a single channel; that is, grayscale. These were provided as a NumPy array. The white horizontal bars indicate the location from which the mandible feature vectors or the one-dimensional data were obtained. The figure was drawn using Matplotlib 3.1.3 (https://matplotlib.org/)[27] under Python 3.7.7 (https://www.python.org). These volunteers agreed to displaying their headshots in this figure.

the labeling and training data, and although it functioned acceptably in most cases, we opted to rescue rare failure cases. Therefore, we manually drew a rectangle that fit each facial contour individually using the image manipulation software GIMP. For all of the images, the positions of the rectangles were subsequently saved as a tab-delimited text file to be processed with OpenCV. Several images were eliminated during this process as they did not meet our requirements. These included images with hands and in which the person was not looking straight ahead. Although we requested the volunteers to remove their glasses, a few volunteers uploaded images with glasses. However, as we felt that the glasses did not affect the classification substantially during the review, we did not eliminate these images. Inclined faces were modified by rotating the entire image to make it upright by using the image manipulation software ImageMagick. The contour positions were obtained from the rotated images. Although very light and dark images existed, the brightness was not modified at all. Finally, 2429 headshot images were obtained from 277 volunteers. For a certain volunteer, only 5 headshots among the 9 uploaded remained, and these were released. Likewise, 6, 7, and 8 uploaded images remained for 4, 11, and 26 volunteers, respectively. For the other volunteers, all 9 images were released, resulting in a total of 2429 images. Examples were obtained from 12 volunteers who provided consent for displaying one of their headshots (Fig. 1).

Comparison with other facial datasets. Table 1 presents the details of well-known facial datasets mentioned in the introduction section, including the number of images and subjects. Datasets collected using web scraping, an image hosting service, and a social network contained more than ten thousand images. Notably, developers downloaded high-resolution photos from image hosting services, and they can refer to the Creative Common licenses for images in Flickr. In the case of FFHQ[16], its website provides a service to verify whether images owned by a specific user account are included and it accepts requests to remove them from the FFHQ dataset. However, when using web scraping, developers cannot obtain agreements from the image owners. Moreover, these images are unconstrained. Constrained headshots can be obtained by a manual collection of data with agreements and informed consent, although the data size would be limited. Low diversity, fairness, and comprehensiveness are additional disadvantages of the manual data collection with agreements. Our dataset, Foxglovetree, has a relatively large number of images and subjects among the constrained datasets (Table 1). FERET was started 30 years ago using 35-mm film cameras with the support of the United States government and includes an exceptionally large number of frontal faces under semi-controlled conditions[7]. As the protection of personal information was underdeveloped at that time, the manner in which consents were obtained from each subject remains unclear. We could not depend on the FERET dataset for practical purposes in the present study in terms of regional population in the collection.

Dataset	Year	Number of image	Number of subjects	Collection	Agreement	Constrained	Notes
FERET	1998	14,126	1199	Manual	Unknown	Semi-	Automatic face recognition
JAFFE	1998	213	10	Manual	Agreement	Yes	Female facial expressions
EBFA	2002	975	13	Manual	Agreement	Yes	Face authentication
LFW	2008	13,233	5749	Web scraping	None	No	Unconstrained face recognition
Adience	2014	26,580	2284	Image hosting service	Creative commons	No	Age and gender prediction
IMDB-WIKI	2015	524,230	Unknown	Web scraping	None	No	Age prediction
AFAD	2016	164,432	Unknown	Social network	None	No	Age prediction
UTKFace	2017	20,000	Unknown	Web scraping	None	No	Labeled by age, gender, and ethnicity
AIST Face2017	2018	136	8	Manual	Informed consent	Yes	Dimensional model of emotions
FFHQ	2019	70,000	14,800	Image hosting service	Creative commons	No	Generative adversarial networks
Foxglovetree	2022	2429	277	CGI upload	Informed consent	Yes	Present study

Table 1. Comparison with facial datasets.

Gender labeling for each volunteer and smile labeling for each image. Because we focused on the appearance rather than biological gender, we did not collect any data regarding gender from the volunteers. Instead, we constructed an in-house voting system to create gender and smile labels. All thumbnails were ordered and displayed using HTML5 to be labeled by the team members. Labeling was carried out by clicking radio buttons and the CGI program subsequently created a voting sheet from each member. All voting sheet sets were finally tallied for the labeling. Therefore, once the subjective voting was completed, the final labeling was objectively completed by computer programs based on majority.

One member was aware that his male friend had provided his images by cross-dressing with make-up. Judging from the appearance in the uploaded images, the member voted for female regarding the male volunteer, without sharing his knowledge with the other members. Because almost all members clicked the female button, the volunteer was labeled as female. This was one of the occurrences during our collecting and labeling processes. However, we did not reveal the image IDs related to these. Among the 277 volunteers, 130 and 147 were labeled as male and female, respectively (Fig. 2).

In addition to gender labeling, we endeavored to label faces with smiling. However, determining whether the person was smiling was very difficult. In hundreds of cases, the team members hesitated to click the radio button. However, the voting system that adopted majority mitigated the manual labeling process. In contrast to the gender labels, smiling was indicated by the numbers of votes obtained, such as 10 and 7 for non-smiling and smiling, respectively. The total number of voters was limited to an odd number of 17 (Fig. 2).

Two-dimensional and one-dimensional data. Based on the rectangles that fit each upright facial contour, we cropped the images and resized them to prepare 359×359 array data for CNN. The faces were decolorized and placed in the center with enough margins (Fig. 1). For example, if the centers are cropped into 299×299 pixels and expanded to three channels, they can be readily used as input data for the Inception V3 network. The headshot data, labels, and the following one-dimensional data were stored and publicly released as NumPy files. The format is described in the method section.

During the data processing, we observed that the gender could simply be distinguished by the existence of long hair around the level of the mouth or jaw in various cases. For the convenience of applying these to other machine learning methodologies, simplified one-dimensional data were also produced, which consisted of one feature vector per volunteer and 75 features derived from horizontal pixel intensities at the jaw level. Bilateral symmetric and neighboring positions were merged to form the 75 features. The merged intensity values were min–max normalized and arranged from the outside to the center (Fig. 1). Hierarchical clustering of the feature vectors was carried out to visualize the values, which suggested the existence of dark pixels around the middle of the array in numerous female samples (Fig. 3). Although principal component analysis could roughly separate the male and female samples, its decomposition ability was far from acceptable (Fig. 3). In addition to these unsupervised learning methods, we carried out linear classification, random forest, and gradient boosting using the feature vectors. The accuracies obtained by the leave-one-out cross-validations were 0.765, 0.791, and 0.801, respectively (Table 2). The top five that contributed to the classification were the 40th, 46th, 42nd, 44th, and 49th features. These were numbered using a 0-based method.

Using the two-dimensional data of the 2429 images with 10-times augmentation, we performed gender classification by deep learning (Fig. 3). In the leave-one-out cross-validation, all headshot data up to 9 from

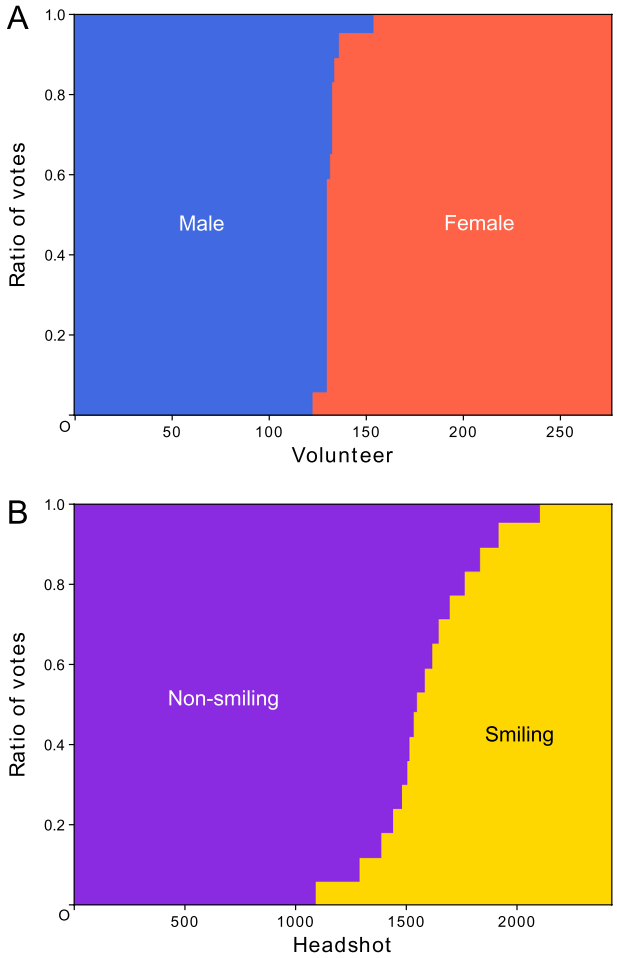

Figure 2. Results of voting for gender and smile labeling. All 277 volunteers and all 2429 headshots were voted on by the team members for the gender and smile labeling, respectively. They were sorted according to the results and renumbered on the horizontal axis. Votes were broken for several volunteers numbered around 130 (**A**) and for the headshots numbered approximately 1400–1800 (**B**). The figure was drawn using Matplotlib 3.1.3 (https://matplotlib.org/) under Python 3.7.7 (https://www.python.org) and modified with Inkscape 1.1 (https://inkscape.org/).

one volunteer were eliminated from the training data and used as test data. This was a leave-one-volunteer-out rather than a leave-one-image-out approach. Compared to other machine learning methods, leave-one-out cross-validation data for CNNs tend to become large. In addition, in this case, the test data should be eliminated from the training data set, as many images, were 10-times augmented from up to 9 headshots for a single volunteer. The size of training and test data varied slightly depending on which volunteer was eliminated because the number of headshots collected was not necessarily 9. By reading the publicly released data, we prepared 277 sets of training and test NumPy files. Although the evidence is not yet conclusive, transfer learning has achieved solid success in deep learning. Because it has been reported that choice of objects did not affect the performance[19], we used the ImageNet data that were pre-trained using various objects, such as birds, bottles, and cards[5]. We adopted Inception V3 with the pre-trained data and trained the network to guess the correct gender for the eliminated test volunteer. This validation process was repeated 277 times, corresponding to each volunteer. For details, refer to Python scripts, prepare_training_and_test_data.py and cross_validation.py, available on GitHub. Out of 277 validations, 272 had successfully predicted the correct gender, *i.e.* 0.982 of volunteer-based accuracy. Because 2342 out of 2429 images were correctly predicted, the image-based accuracy was 0.9642. When the same augmented data were applied to random forest, the image-based accuracy was 0.8600 (Table 2).

Finally, we carried out smiling recognition with the same released data. Because we prepared smiling labels by the number of votes, users have to set thresholds to select smiling and non-smiling headshots. When we chose the smiling data that won more than 80% of the vote, 668 images were obtained out of the 2429. To maintain balance, we additionally chose 668 data that received exclusively non-smiling votes. Unlike gender recognition, smiling recognition should be an image-based and not a volunteer-validation. Because the number of images was larger than that of the volunteers, it was difficult to perform the leave-one-out cross-validation. With 10-times augmentation and identical CNN, we performed tenfold cross-validation, yielding an accuracy of 0.9303 (Table 2). These scripts are also available on GitHub for replication and to explore these results.

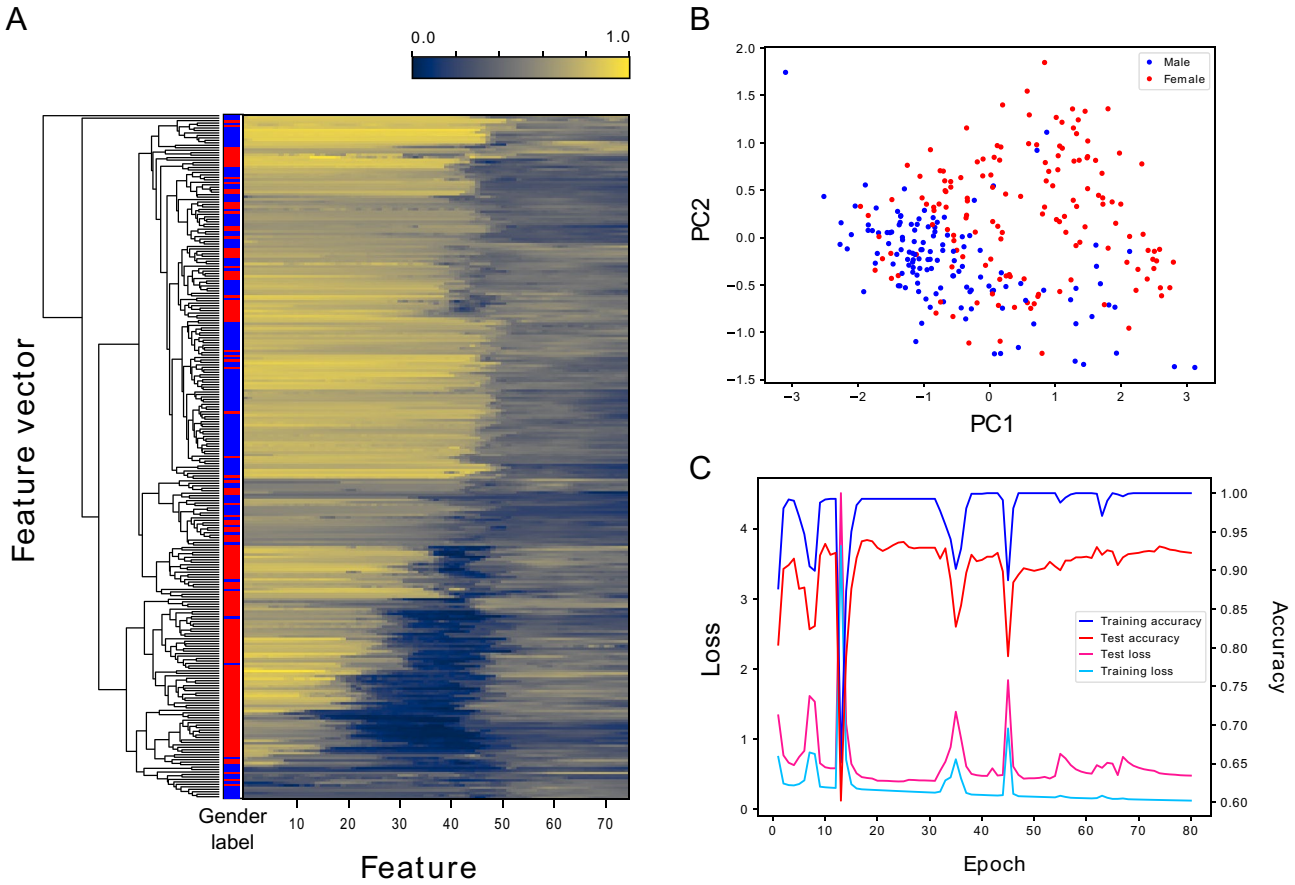

Figure 3. Hierarchical clustering, principal component analysis, and deep learning. Hierarchical clustering was performed for the mandible feature vectors, which consisted of 277 samples and 75 features (**A**). The darker color corresponds to the darker parts in the headshot image. Although genders were not separated effectively by this unsupervised method, female samples tended to exhibit darker parts around the halfway point. Principal component analysis was performed for the feature vectors (**B**). The blue and red dots represent 130 male and 147 female samples, respectively. Although they preferentially occupy the bottom left and top right corners, respectively, the two gender groups could not be clearly discerned by this unsupervised method. Learning curves of the CNN based on the Inception V3 is shown (**C**). Since the leave-one-out cross-validation with 10-times augmentation could achieve high accuracy in a single epoch, learning curves of a twofold cross-validation without data augmentation are shown. Odd-numbered and even-numbered volunteer samples were used as training and test data. Exemplary curves were observed during the first 5 epochs in this case. Cyan, pink, blue, and red lines indicate training loss, test loss, training accuracy, and test accuracy, respectively. The figure was drawn using Matplotlib 3.1.3 (https://matplotlib.org/) under Python 3.7.7 (https://www.python.org) and modified with Inkscape 1.1 (https://inkscape.org/).

Label	Method	Number of headshots	Pixels per sample	Augmentation	Cross-vallidation	Accuracy	F-score
Female	Jubatus	277	75	1×	Leave-one-volunteer-out	0.765	0.767
Female	Random forest	277	75	1×	Leave-one-volunteer-out	0.791	0.803
Female	LightGBM	277	75	1×	Leave-one-volunteer-out	0.801	0.812
Female	Random forest	2429	299 × 299	10×	Leave-one-volunteer-out	0.8600	0.8671
Female	CNN	2429	299 × 299	10×	Leave-one-volunteer-out	0.9642	0.9621
Smilling	CNN	668 × 2	299 × 299	10×	10-fold	0.9303	0.9297

Table 2. Summary of the presented classification models.

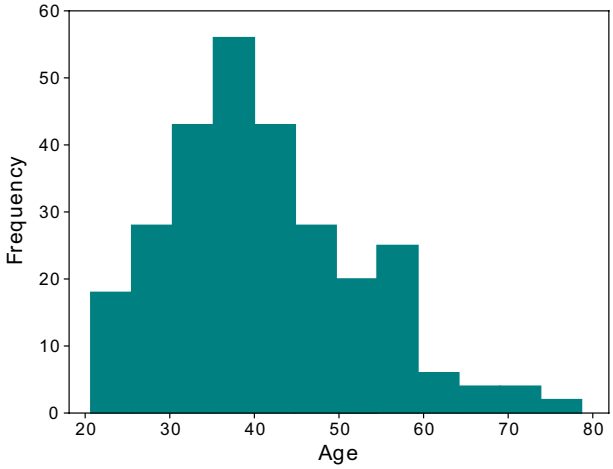

Figure 4. Distribution of volunteer age. Because age was calculated from the year of birth on the agreement form, there could be ± 1 year error. The minimum, median, and maximum ages were 21, 40, and 79 years old, respectively. The figure was drawn using Matplotlib 3.1.3 (https://matplotlib.org/) under Python 3.7.7 (https://www.python.org) and modified with Inkscape 1.1 (https://inkscape.org/).

Discussion

Over the course of approximately one year, we recruited more than 200 volunteers online. The campaign was not straightforward. We initially anticipated that young people would be interested and advertised the initiative through several social networking services. However, contrary to our expectations, the number of volunteers barely increased until we held recruiting seminars repeatedly. Calling out to the employees at our institute also contributed to the final quantity of the collected data. The online collection system required each volunteer to validate his or her year of birth to confirm that they were no younger than 20 years old. Although we neither used the data for deep learning nor released the data, the percentage of volunteers that were less than 30 years old was 14.8%. It is likely that older people tended to be more lenient in providing their own photographs (Fig. 4).

The campaign was challenging owing to the current trend of protecting personal information. One of the contributions of the present study is the demonstration that all photographs were collected according to proper procedures. Prior to collecting the data, we consulted with members of the ethics review committee several times to determine what was and was not allowed. Although our proposal using the online system was finally approved, we had to compromise by limiting the types of collected data. It is unfortunate that we could not release colorized or three-channel data. However, the data quantity would have been much lower if we had not decolorized the images. Unlike in web scraping, all volunteers were requested to take a front-facing posture. The quality thus obtained helped reduce the amount of training data needed. We have demonstrated that using all these volunteer data is sufficient to predict their gender and smiling using CNN. Many congenital disorders such as Down syndrome, Noonan syndrome, and Williams syndrome can be diagnosed to some extent from their characteristic facial features by experienced doctors[20]. Because most of these are rare diseases, collecting data would be a demanding task; however, our results encourage the medical researchers who are planning to build new AI systems to proceed with such tasks. Constrained-posture photographs and natural, unconstrained ones are valuable for promoting further research. We continue to seek more sophisticated means of collecting even more constrained facial data with informed consent using other technologies, such as blockchain.

Although requesting 9 photographs appeared to place a heavy burden on volunteers, many images of a single person aided us in creating labels unexpectedly. This was applicable to both the gender and smile labeling. In certain female cases, it was difficult to draw conclusions using a single photograph. Reviewing several facial expressions may play a role in the recognition of a person. As the selection of smiling faces was rather subjective, many images were certainly necessary. Because most photographs were taken in "selfie" modes, natural smiling did not appear frequently in our collections. We initially thought that constrained headshots were not suitable for smile labeling; however smiling recognition worked unexpectedly well, probably due to the labeling system by vote.

Several studies have concluded that face recognition is more accurate in men than in women[17,21]. Accordingly, our CNN committed errors in 2 male and 3 female cases. All 3 females had short hair; one of the 2 males had long hair. Compared to unconstrained datasets, our collection was relatively small. However, as a constrained headshot dataset with informed consent and hand-annotations, the scale was unprecedented. The resolution of each image was much higher than that of the images obtained by web scraping. Obtaining several images—up to 9—from each individual was also a significant advantage. We could straightforwardly observe the importance of hair length, which was also supported by the one-dimensional mandible vectors. Despite the fact that our collection system was initially built for educational purposes, it appeared to offer the possibility of identifying unrecognized biometric traits in the faces.

Conclusion

By developing an online system, constrained headshots with informed consent were obtained from 277 volunteers. In total, 2429 images were hand-annotated and publicly released. While web scraping is a convenient way to collect face images, a high accuracy of gender and smiling recognition using these relatively small data suggests the importance of quantity and quality of constrained headshots.

Methods

Ethics approval and protection of personal information. All methods were performed in accordance with the relevant national guidelines and regulations. The present study was approved by the Institutional Review Board at the National Center for Child Health and Development, approval number 2045 on January 10, 2019. As described in the result section, informed consent was obtained online, which was also approved by the review board. Written informed consent for publication of identifying images was obtained from the 12 volunteers appeared in Fig. 1. They all agreed to displaying one of their headshot images in an online open-access publication.

Hardware and software. We shared NEC Express5800/R120d-2E in addition to individual personal Linux, macOS, or Windows computers. We used Hitachi HA8000/RS210, HPE ProLiant DL360, and Hitachi SR24000/DL1 for the high-intensity calculations. For data collection, data release, and demonstration of our own application, we used hosting services with global IP addresses. We made use of numerous online services, including FC2 Wiki, Slack, Facebook, GitHub, and Google Docs, without any payments. For the most part, we used CentOS 7.6, but we attempted various Linux distributions, from Ubuntu and openSUSE to Windows Subsystem for Linux. We exclusively employed free software or open-source software as far as possible, including Anaconda, TensorFlow[22], Keras[23], OpenCV, GIMP, ImageMagick, Jubatus[24], Scikit-learn[25], and LightGBM[26].

Headshot data and labels. The shape of the released images was a 359×359 pixel square, which was reshaped into a 128,881-dimensional vector row by row. To ensure the equality of the volunteers, the vectors were shuffled and concatenated into an array saved as a NumPy file. The presented order could be retrieved from the bundled label data, which were provided as another NumPy file. The file consisted of five columns: volunteer ID, photograph number for a given volunteer, female label, non-smile label, and smile label. The volunteer IDs were serial numbers from 1 to 277. The photograph numbers ranged from 1 to 9, but there were missing numbers for certain volunteers. If the volunteer was considered a female by our decision rule, the column was labeled as 1. Non-smiling and smiling labels were indicated by numbers of votes.

Mandible feature vectors. In addition to the headshot data, a simplified dataset was provided. The data were derived from horizontal pixel intensities around the level of the mouth or jaw, in the 267th, 268th, and 269th rows from the top, excluding 30 pixels on both sides. These were numbered with a 0-based method. The 6 neighboring pixels were merged, except for the 9 center pixels. Furthermore, bilateral symmetric positions were merged to form the 75 features or explanatory variables. The merged intensity values were min–max normalized, from 0.0 to 1.0, and arranged from the outside to the center around the mouth.

Unsupervised and supervised machine learning. Hierarchical clustering was carried out using the Euclidean distance and average linkages. Clustering and drawing of the heatmap and dendrogram were achieved with the SciPy library 1.1.0, https://scipy.org/. Principal component analysis was carried out using the Scikit-learn library 0.21.2, https://scikit-learn.org/. Linear classification was performed using Jubatus 1.1.1, http://jubat.us/, with the AROW algorithm, a regularization weight of 1.0, and 100 epochs. Random forest was applied using the Scikit-learn library with a maximum depth of 32, 16 trees, and a random state of 32. Gradient boosting was carried out using LightGBM 2.3.0, https://github.com/microsoft/LightGBM. Binary log loss classification was selected with 38 boosting rounds. If a predicted value was no less than 0.5, the sample was inferred as female. Inception V3 with the ImageNet pre-trained data yielded the optimal results for the CNN gender classification. The network was implemented in Keras 2.2.4, https://keras.io/, with TensorFlow 1.12.0, https://www.tensorflow.org/, as a backend. We added three fully connected layers and a sigmoid layer for the binary classification. We set the 225th and latter layers to be trainable. The training data were augmented 10 times using Keras ImageDataGenerator with a rotation range of 6.0, width shift range of 0.06, height shift range of 0.06, shear range of 2, zoom range of 0.08, and horizontal flip on. The batch size and number of epochs were 64 and 20, respectively. When the two-dimensional data were applied to random forest, they were flattened resulting in 89,401 features. Gender and smiling recognitions were evaluated by leave-one-out and tenfold cross-validation, respectively. Our Python scripts are available at https://github.com/glires/Foxglovetree/. Those used for Jubatus, random forest, LightGBM, and CNN are saved in correspondingly named subdirectories in the analysis directory, respectively.

Availability of data and programs. The image data and labels, namely the Foxglovetree dataset, are publicly available from the following sites, https://aihospital.ncchd.go.jp/foxglovetree/data/ or https://github.com/glires/Foxglovetree. In appreciating the generosity of many volunteers, we allow neither distribution nor release of any image formats, such as JPEG files, derived from our data. The scripts used to process the data and to perform machine learning are also available at https://github.com/glires/Foxglovetree or our portal site, https://aihospital.ncchd.go.jp/.

Received: 8 July 2020; Accepted: 9 February 2022
Published online: 08 March 2022

References

1. Zahangir Alom, M. *et al.* The history began from AlexNet: a comprehensive survey on deep learning approaches. Preprint at https://arxiv.org/abs/1803.01164 (2018).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1*; 1097–1105 (Curran Associates Inc., 2012).
3. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
5. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. Preprint at https://arxiv.org/abs/1409.0575 (2014).
6. Golomb, B. A., Lawrence, D. T. & Sejnowski, T. J. SEXNET: a neural network identifies sex from human faces. In: Lippmann, R., Moody, J. E. & Touretzky, D. S., editors. NIPS. conf/nips/GolombLS90: Morgan Kaufmann; 572–579 (1990).
7. Phillips, P. J., Wechsler, H., Huang, J. & Rauss, P. J. The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* **16**, 295–306 (1998).
8. Lyons, M., Kamachi, M. & Gyoba, J. The Japanese Female Facial Expression (JAFFE) Dataset. Zenodo at https://doi.org/10.5281/zenodo.3451524 (1998)
9. Liu, X., Chen, T., & Kumar, B. V. K. V. On modeling variations for face authentication. In *International Conference on Automatic Face and Gesture Recognition*. 369–374 (2002).
10. Huang, G. B., Mattar, M., Berg, T. & Learned-Miller, E. 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*; 2008–10 (2008).
11. Eidinger, E., Enbar, R. & Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.* **9**, 2170–2179 (2014).
12. Rothe, R., Timofte, R. & Van Gool, L. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*; 252–257 (2015).
13. Niu, Z., Zhou, M., Wang, Wang, L., Gao, X. & Hua, G. Ordinal regression with a multiple output CNN for age estimation. CVPR 4920–4928 (2016).
14. Zhang, Z., Song, Y. & Qi, H. Age progression/regression by conditional adversarial autoencoder. Preprint at https://arxiv.org/abs/1702.08423 (2017).
15. Fujimura, T. & Umemura, H. Development and validation of a facial expression database based on the dimensional and categorical model of emotions. *Cogn. Emot.* **32**, 1663–1670 (2018).
16. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. Preprint at https://arxiv.org/abs/1812.04948 (2019).
17. Buolamwini, J. & Gebru, T. Gender Shades: intersectional accuracy disparities in commercial gender classification. In: Sorelle, A. F. & Christo, W., (eds.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency; Proceedings of Machine Learning Research: PMLR*; 77–91 (2018).
18. Bradski, G. The OpenCV library. *Dr. Dobb's J. Softw. Tools* (2000).
19. Huh, M., Agrawal, P. & Efros, A. A. What makes ImageNet good for transfer learning? Preprint at https://arxiv.org/abs/1608.08614 (2016).
20. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
21. Albiero, V. *et al.* Analysis of gender inequality in face recognition accuracy. Preprint at https://arxiv.org/abs/2002.00065 (2020).
22. Abadi, M. *et al.* TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org (2015).
23. Chollet, F. *Deep Learning with Python* (Mamming, 2017).
24. Hido, S., Tokui, S. & Oda, S. Jubatus: an open source platform for distributed online machine learning. NIPS 2013 Workshop on Big Learning (2013).
25. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
26. Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *31st Conference on Neural Information Processing Systems* (2017).
27. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgements

Our small group activity was originally encouraged by the director of the research institute, Dr. Yoichi Matsubara, and subsequently by the NCCHD AI hospital project, particularly by Dr. Yasuo Kiryu. Ms. Kyoko Hagino provided us with constructive suggestions regarding the protection of personal information. Although we did not receive any financial support for the activity itself, the publication cost was supported by the AI hospital project, officially, "Innovative AI Hospital System", Cross-ministerial Strategic Innovation Promotion Program (SIP), Council for Science, Technology and Innovation (CSTI). The National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN) is its funding agency. It was also supported by Takeda Science Foundation. We did not purchase any hardware for this study. However, several departments and individual employees generously provided used computers. Our main machine, NEC Express5800/R120d-2E, was provided by BioBank, which is directed by Dr. Kenichiro Hata. We also used the Hitachi HA8000/RS210 and HPE ProLiant DL360 clusters at the Center for Regenerative Medicine, which is directed by Dr. Akihiro Umezawa, Hitachi SR24000/DL1 P100 purchased by Dr. Yoichi Matsubara, and the same model with V100 purchased by the AI hospital project. Finally, we sincerely appreciate all of the volunteers for providing their headshot image data.

Author contributions

K.O. conceived and designed the study. Y.I. applied for ethics review. S.A., M.H., H.U., A.U., M.I., Y.I., M.T., T.J., M.S., Y.O., F.H., H.O., S.N., K.K., M.K., K.M., K.T., and K.O. collected, curated, annotated, and analyzed the data and discussed the results. K.O. compiled the release data. M.T., K.T., and K.O. drafted the manuscript. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.