Face Verification with Challenging Imposters and Diversified Demographics

Adrian Popescu¹, Liviu-Daniel Ştefan², Jérôme Deshayes-Chossart¹, Bogdan Ionescu²
¹Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
²University Politehnica of Bucharest, Romania
{adrian.popescu,jerome.deshayes-chossart}@cea.fr, {liviu_daniel.stefan,bogdan.ionescu}@upb.ro

Abstract

Face verification aims to distinguish between genuine and impostor pairs of faces, which include the same or different identities, respectively. The performance reported in recent years gives the impression that the task is practically solved. Here, we revisit the problem and argue that existing evaluation datasets were built using two oversimplifying design choices. First, the usual identity selection to form impostor pairs is not challenging enough because, in practice, verification is needed to detect challenging imposters. Second, the underlying demographics of existing datasets are often insufficient to account for the wide diversity of facial characteristics of people from across the world. To mitigate these limitations, we introduce the FaVIC12D dataset. Imposter pairs are challenging because they include visually similar faces selected from a large pool of demographically diversified identities. The dataset also includes metadata related to gender, country and age to facilitate fine-grained analysis of results. FaVIC12D is generated from freely distributable resources. Experiments with state-of-the-art deep models that provide nearly 100% performance on existing datasets show a significant performance drop for FaVIC12D, confirming our starting hypothesis. Equally important, we analyze legal and ethical challenges which appeared in recent years and hindered the development of face analysis research. We introduce a series of design choices which address these challenges and make the dataset constitution and usage more sustainable and fairer. FaVIC12D is available at https://github.com/AIMultimediaLab/FaVIC12D-Face-Verification-with-Challenging-Imposters-and-Diversified-Demographics.

1. Introduction

Face verification (FV) is deployed in applications such as biometrics [8, 20], social media information structuring [16] or classical media archive organization [5]. Performance has strongly progressed in recent years due to the introduction of deep learning techniques [23, 40]. Results reported for public datasets collected from heterogeneous sources, such as LFW [21], IJB-C [24], MegaFace [19] or TrillionPairs [3], are getting very close to 100% accuracy.

We find that such performance is misleading due to two design choices made when building existing face verification datasets. Following [27] and [36], we hypothesize that the usual random selection of identities for impostor pairs makes the verification process too easy. Challenging impostor pairs in FaVIC12D are created by using a deep face recognition model to get visually similar imposters. Then, a manual verification is done to ensure that the two identities are actually different. Challenging genuine pairs are equally interesting to test the limits of verification systems. They are created by asking annotators to select images of the same identity which are visually different.

A second problem is the scale of the pool of imposters and the fairness of the face verification process. Imposter pool size is not central for the random selection of imposters if the pool is not too low. This parameter becomes important for challenging imposters since the dataset should contain faces with similar demographics for each target identity. Diversified and balanced demographic distribution of imposters is equally central to ensure a fair verification process [12]. If impostor demographics is imbalanced, a bias that is inverse compared to face recognition appears. A larger imposter pool from a given demographic split makes the verification process more challenging. The more imposters there are, the larger the chances of finding a similar one for a target identity. Note that scale was addressed in MegaFace [19] or DiF [25], while fairness was only partially addressed in smaller datasets such as FairFace [18] or BFW [29]. FaVIC12D addresses these technical challenges. Imposter pairs are challenging because they include visually similar faces selected from a large pool of diversified identities. The dataset also includes demographic metadata to facilitate fine-grained analysis of results.

Legal and ethical aspects are central in face verification because the task deals with sensitive information related to subjects’ identity. Ignoring or minimizing such aspects probably contributed to the withdrawal of datasets [34],
such as MS-CELEB1M \[14\], MegaFace \[19\] and DiF \[25\]. Legal criteria were considered during the constitution of FaVCl2D, notably in terms of enforcing copyright, personal data protection, and image rights. Reuse requirements are tackled by exploiting only resources which are released under suitable licenses. Data protection regulations are different across the world. A recent comparison between the US, the EU and China’s approaches to data protection \[26\] concludes that the highest level of protection is offered by EU’s General Data Protection Regulation (GDPR) \[2\]. This is particularly the case for sensitive data, such as faces. Consequently, the latter regulation should be taken into account when building face-related datasets. Ethical aspects are also key in order for face verification and recognition to be developed in a socially acceptable way. Some of the concerns that have been voiced address:

- Unlawful and/or unethical dataset collection \[15\] with problems such as: disclosure of the names of the persons, inclusion of copyrighted images and insufficient handling of consent, especially for children.
- Bias against demographic categories, such as gender, age or origin \[6, 18, 29, 36\]. This is notably an effect of a strongly imbalanced collection of the face recognition datasets used to create deep models.
- Banning face recognition for law enforcement \[11\] and surveillance \[10, 13\]. These initiatives are part of a larger debate regarding risks related to AI technologies. The raised objections should be carefully considered both from technical and legal perspectives to improve the acceptability of face-related technologies.

Our work is informed by these concerns and effort is devoted to addressing them. During dataset design, we put focus on: (1) compliance with legal requirements, (2) data minimization by storing only information necessary to fulfill the task, and (3) reducing demographic imbalance to ensure a fair analysis of demographic segments. Note that FaVCl2D is built for face verification and, given the reduced number of faces per identity and anonymization of its identities, would be of no use to build recognition models directly. FaVCl2D could be used during the construction of future recognition datasets to better calibrate them in terms of demographic representativity. It is also noteworthy that, while the focus is put on security-related applications, face verification is useful in a range of other contexts. For instance, is increasingly used to organize large multimedia collections and thus improve access to their content \[16, 5\].

2. Analysis of Face Verification Datasets

Verification performance reported on widely-used datasets, such as LFW \[21\], IJB-C \[24\], MegaFace \[19\], is close to 100%. Such performance suggests that the task is solved or nearly so. However, the authors of \[36\] show that if challenging imposter pairs are introduced in face verification, performance drops significantly. This design choice makes face verification more realistic, and we build upon it in our work. Given the sensitiveness of the task, much attention was given recently to different biases. The influence of gender, age, and ethnic origin was discussed, among others, in \[6, 18, 29, 36\] and should be carefully considered. With \[32\], we note that non-demographic factors also introduce biases in face verification and deserve investigation.

2.1. Face Verification Design Criteria

We analyze datasets taking into account technical, legal and ethical aspects. We propose the following technical characteristics for a sound design of verification datasets:

- \(T_1\): Development should be guided by real-life usage of this technology. The inclusion of hard genuine and imposter pairs \((T_{1.1} \text{ and } T_{1.2})\), respectively, is important to challenge the evaluated deep representations.
- \(T_2\): The number of identities \((T_{2.1})\) and of images \((T_{2.2})\) should be large enough to approximate large-scale face verification systems.
- \(T_3\): A balanced spread of identities in terms of demographic factors such as gender \((T_{3.1})\), geographic origin \((T_{3.2})\) and age \((T_{3.3})\) should be achieved.
- \(T_4\): Datasets should include identities which are representative for the general population, notable\(^1\) or not.

We devise the following legal or ethical characteristics:

- \(L_1\): FV datasets should be built on top of resources whose licenses allow reuse and modification. The use of other raw data was shown to be problematic in the long run and led to the withdrawal of some datasets.
- \(L_2\): Compliance with data protection regulations is needed for a lawful distribution and usage of the dataset. A comparative analysis of data protection laws \[26\] concludes that the EU’s GDPR offers the highest level of protection for sensitive data. Art. 9 forbids the processing of such data, with exceptions for research in art 9.2(j) if proportionality with the aim pursued is established as described in art. 89.
- \(L_3\): Compliance with other privacy laws, particularly with the image rights applicable in countries such as Canada, Belgium, France, or Spain. This right basically forbids the distribution of images that include recognizable faces. An exception is made for public figures when they appear in a professional capacity.
- \(L_4\): Sustainability of the dataset is crucial for its future use. Some of the datasets were withdrawn because they generated strong debates about their adherence to legal and ethical standards \[15\].

The simultaneous optimization of all technical and legal criteria is difficult, if not impossible. For instance, it is de-

\(^{1}\)https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)
Tabelle 1. Überblick über Face-Verifikationsdatensätze aufgrund von wünschenswerten Charakteristika. Gute (T1,1) und Impostorpaare (T1,2) sind entweder zufällig oder herausgefiltert. Werte sind für einzigartige IDs (T2,1) und für Bilder (T2,2), mit Störerpaaren, die verwendet wurden. Geburtsjahr (T3,1) ist die Proportion von männlichen/Weiblichen Bildern. Ursprung (T3,2) wird durch das Anwenden von Geschlecht, Hautfarbe oder Länder berücksichtigt. Origin-related Unbalance wird erwähnt. Alter (T3,3) wird manuell oder automatisch bestimmt oder tatsächliche, wenn die Aufnahme gemacht wurde. ID-Typ (T3,3) kann entweder bekannt oder unbekannt sein. Legal und ethische Charakteristika (L1 – L4) sind mit Ja/Nein beschrieben. NK steht für „nicht bekannt“ wenn Information nicht verfügbar war.

| Dataset            | LFW [21] | YTF [37] | IJB – C [24] | MegaFace Pairs [19] | Trillion Pairs [18] | DIoF [35] | FaceTime [19] | RFW [36] | IJB – C ext [1] | FaVC12D (proposed) |
|--------------------|----------|----------|--------------|---------------------|--------------------|----------|---------------|----------|----------------|------------------|
| T1,1 genuine       | random   | random   | random       | random              | random             | random   | random        | random   | random         | random           |
| T1,1 impostor      | random   | random   | random       | random              | random             | random   | random        | random   | random         | random           |
| T2,1 unique IDs    | 5.749    | 1.595    | 3.531        | 530 +distractors    | 5.749 +distractors | NK       | NK            | NK       | 3.000          | 6.139            |
| T2,1 total images  | 13,233   | 621,400  | 31,304       | 1,000,000           | 1,000,000          | 108,301  | 1,000         | 152,917  | 64,879         |
| T1,1 gender (F/M)  | 22.5/77.5| 37.3/62.7| 41.1/58.9    | 22.5/77.5           | 48/52              | 30/50    | 35/65         | 38,5/61.5| 44/56          |
| T2,1 origin        | race imbal. | race imbal. | race imbal. | race imbal. | race imbal. | skim color imbal. | race bal. | race bal. | race skim color imbal. | country imbal. |
| T3,1 age           | NK       | NK       | NK           | estimated           | estimated          | estimated | estimated     | estimated | actual          | actual           |
| T1,1 ID type       | notable  | notable  | notable      | any                 | any                | any      | any           | any      | any            | any              |
| L1 reusability     | no       | no       | yes          | yes                 | no                 | yes      | yes           | yes      | yes            | yes              |
| L2 privacy         | no       | no       | no           | no                  | no                 | no       | no            | no       | yes            | yes              |
| L3 image rights    | yes      | no       | yes          | no                  | no                 | no       | no            | yes      | yes            | yes              |
| L4 sustainability   | yes      | yes      | yes          | yes                 | no                 | yes      | yes           | yes      | yes            | yes              |

Es ist wünschenswert, eine große Anzahl an Ids im Datensatz (T2). Jedoch, die Verfügbarkeit von Identitäten ist variabel für Faktoren wie Geschlecht und geografische Herkunft [35]. Dieser Faktor beeinflusst die Erreichbarkeit von unterschiedlichen regionalem (T3) (T1,3). Die Verfügbarkeit aller Anwendbaren Images (T2) ist größer als die Verfügbarkeit von Images (L1) und das Verwenden von Copyrighted Images ist riskant. Wir haben den besten Versuch, so viele Kriterien wie möglich zu erfüllen. In Tabelle 1, analysieren wir neun existierenden FV-Datensätze und FaVC12D. Wir integrieren technische und rechtliche/ethische Charakteristika, die wichtig sind für Nachhaltigkeit und unabhängige Evaluierung von Datensätzen. Negativere Effekte haben eine negative Auswirkung auf die öffentliche Bewertung der Face-Verifikation und schließlich zwingen uns, die Entwicklung zu unterbrechen.

2. Analyse der technischen Kriterien

Die ersten zwei Charakteristika, die in Tabelle 1 referieren, die Guten (T1,1) und Impostorpaare (T1,2), sind entweder zufällig oder herausgefiltert. Wenn wir erwähnt haben, diese Kriterien sind wichtig, um eine reale Bewertung der Face-Verifikation zu gewährleisten. Die Impostorpaare sind wertvoll, weil die Verifikation in schwierigen Bedingungen zeigt, wie robust der Prozess ist. Wir schätzen die sehr hohen Leistungen auf der Basis von Performanceberichten aus bestehenden Datensätzen wie LFW [21], IJB-C [24] oder IJB – C_ext [1], wir schätzen, dass die Verwendung von herausgefilterten genuine-paaren ist, um die getesteten Features besser zu vergleichen. Die Verwendung von herausgefilterten genuine-paaren ist eine deception-oriented-Scenarios, in denen jeder Impostor Paar versucht, das Verifikationssystem zu überwinden. Wir vergleichen herausgefiltertes und random Impostorpaare, um die Unterschiede zwischen ihnen zu messen. Note that only RFW [36] and FaVC12D include challenging genuine-impostor pairs, which are created using visual similarities between IDs. A related problem is that of the amount of curation applied to the included faces. Biometric verification works with a curated target (for instance, an ID photo) and a non-curated query image. Other scenarios, such as verification in media archives or in social media, require verification in absence of curation since images in a pair come from uncontrolled sources. Similarly to all recent datasets analyzed here, FaVC12D includes non-curated images. The dataset is thus best fitted for usage for the second type of scenarios which gained a lot of traction.

The number of unique IDs (T3,1) used as a probe is another important criterion insofar it enables a thorough evaluation. Existing datasets include thousands of identities to form genuine pairs. FaVC12D has the highest number of unique IDs (12,468) among the datasets for which this number is known. MegaFace [19] and Trillion Pairs [3] also include distractors to form a large number of diversifed impostor pairs, as does FaVC12D. However, since selection is random for MegaFace and Trillion Pairs, the utility of using many distractors and a very large number of pairs is questionable. It would have been possible to further expand FaVC12D but at the price of increasing demographic imbalance, which is one of the main criticisms associated with existing datasets [6, 18, 29]. The limitation comes from the number of notable persons belonging to underrepresented demographic segments, such as women from African countries. This limitation is particularly strong when searching for pairs of representative and reusable images. Ongoing projects such as “Wiki loves women”² aim to reduce the demographic imbalance in Wikipedia. We will later release revised versions of the dataset to reflect such changes. The effect of the number of unique impostor IDs and the number of unique IDs is evaluated in the experimental section.

The total number of images (T2,2) varies a lot across datasets. Megasface and Trillion Pairs include the highest number of images since they exploit a very large number

²https://en.wikipedia.org/wiki/Wikipedia:WikiProject_WikiLovesWomen
of distractors. \textit{YTF} \cite{37} was built from videos and it was easy to generate a large number of frames. \textit{FaVC12D} has fewer images because we choose to use only two images per ID to form genuine pairs and one image for imposter IDs. This choice is made to enforce data minimization (see $L_c$).

Gender distribution ($T_{3.1}$) is a known problem in face verification and recognition \cite{6}. LFW, IJB-C, MegaFace, Trillion Pairs, RFW, and IJB – $C_{ext}$ are strongly imbalanced. Efforts toward gender parity were made for DiF and FairFace. In \textit{FaVC12D}, we wanted to diversify geographic spread and were able to achieve balance for Asia, America and Europe. Unfortunately, reusable data was scarce for Africa. We chose to match the number of IDs of African origins with those from other regions at the expense of strict gender parity for this region.

Origin ($T_{3.2}$) is another highly sensitive criterion for which it is difficult to propose an objective and uncontroversial segmentation. A majority of datasets use the notion of race to group people on this criterion. However, the concept of race is, to say the least, controversial \cite{22}. Its use has also strongly contributed to the controversies which led to the withdrawal of face recognition datasets \cite{34}. Following \cite{1} and \cite{25}, we decided to discard it in \textit{FaVC12D}.

Skin color \cite{25} is more objective than race but we decided against its use since: (1) the same skin tone can characterize people of different origins or polyethnic combinations, (2) it can greatly vary due to the conditions in which a photo was taken and, more anecdotally, to tanning. Instead, we group people by their country of origin. We acknowledge that this segmentation is equally imperfect because: (1) many countries have borders that do not correspond one-to-one to ethnic groups, (2) some people have multiple citizenship, and (3) a large part of people are polyethnic. However, this criterion is objective and less likely to lead to controversies than race or its approximation via skin color.

Age is the third important demographic criterion ($T_{3.3}$) which should be tested in face verification. Some datasets have no age-related information, e.g., LFW, YTF, MegaFace, while others estimate it automatically, e.g., DiF, FairFace, RFW. Following \cite{1}, we decided to use the actual age of the persons when they were photographed. Since not all images have time-related metadata associated to them, we build subset of \textit{FaVC12D} to estimate the effect of age.

The type of IDs included ($T_4$) is another important criterion. While appealing, the use of faces of “common” people is legally and ethically difficult. Explicit consent would be needed from each person included for GDPR-compliance. Such a constraint is highly impractical at a large scale but it can be waived for notable persons (see $L_2$ and $L_3$ below). Ethical challenges are equally strong because the inclusion of identities other than notable led to the withdrawal of DiF \cite{25} and MegaFace \cite{19}. The type of included IDs should therefore be carefully considered during dataset design to ensure sustainable exploitation. One interesting question that appears is whether the facial characteristics of notable persons are different from those of other people and thus affect the representativity of verification results. Compared to existing datasets, the proposed demographic diversification reduces the probability for the retained sample faces to be different from the general population.

### 2.3. Analysis of Legal and Ethical Criteria

The type of license associated with the images ($L_1$) is a first important legal criterion. The inclusion of copyrighted images contributed to the withdrawal of MS-Celeb1M \cite{34} but datasets derived from it, including Trillion Pairs and RFW, are still distributed. We note that a majority of analyzed datasets were created from reusable content. Access to \textit{FaVC12D} will be granted only after the signature of a contract which will specify the rights and obligations of the users, notably concerning exclusive use for scientific research purposes.

Data protection ($L_2$) should be enforced when working with sensitive data such as faces. Notably, data minimization stipulates that resources should only include information needed to carry out a technical task in a sound way. Redundant information is included in most analyzed datasets. For instance, not all ID images from LFW are used to create pairs. \textit{FaVC12D} complies with GDPR requirements on sensitive data because it is designed for research purposes (\textit{art. 9(j)}). It instantiates data minimization (\textit{art. 89.1}) by: (1) storing two images per identity which is the minimum quantity needed to perform verification, (2) removing the names of the persons from the dataset, (3) ensuring that all demographic segments include a sufficient number of identities. The following data-related rights are implemented to comply with \textit{art. 89.2}: right to access (\textit{art. 15}), right to rectification (\textit{art. 16}), right to restriction of processing (\textit{art. 18}) and right to object (\textit{art. 21}). A Web form through which any of the persons included in the dataset can require the expression of their rights will be made available. The proposed data protection measures establish proportionality between the proposed usage of data and the rights of the data subjects. To our knowledge, there is no publicly available legal analysis of GDPR compliance for face-related datasets. However, the proposed measures are in line with the recommendations made for the processing of genetic data \cite{31}, which also fall in the sensitive category defined by GDPR in \textit{art. 9}.

Image rights ($L_3$) are respected if only notable people in public situations are included. The repurposing of large datasets, such as YFCC100M \cite{33}, led to the creation of MegaFace, DiF and FairFace. While technically tempting, it is legally challenging in a range of countries. For instance, the inclusion of children’s faces in MegaFace, underlined in \cite{15}, is problematic. To be on the safe side, no children’s
faces were used in FaVCI2D. The use of Wikipedia images minimizes the risk of including photos taken in private contexts, but a manual verification of the photos was still performed to exclude this risk.

Last but not least, sustainability (L₄) is a core criterion for dataset usefulness. MegaFace and DiF were already withdrawn and Trillion Pairs, FairFace or RFW might be next due to public pressure. The set of measures presented above should ensure long-term availability of FaVCI2D.

3. Proposed Dataset

3.1. Identity selection and processing

The first step of the data collection is to create a list of diversified identities with metadata associated to them. Similarly to existing face recognition or verification datasets [7, 14, 21], FaVCI2D includes notable people. Differently from them, we aim to create a demographically diversified dataset by systematically exploiting metadata.

Identities are selected using a series of filters associated to demographic factors. The first filter is related to age. Only articles which are categorized under “YYYY births” in Wikipedia are kept, with 1920 ≤ YYYY ≤ 2000. The first bound is set to cover a wide variety of ages, while the second is used to avoid including images of children.

Geographical spread is the second filter used. Wikipedia entries are biased toward Western European, North American and other populous countries such as India and Japan [28]. This bias is reduced by imposing a maximum number of 10,000 entries per country for inclusion in FaVCI2D. Country names and associated demonyms for 183 countries are first searched in the article categories. If several of them occur, the one with the maximum count in the entire article is retained. We checked country attributions for 500 identities and they were 99.4% accurate.

Gender balance is targeted at the regional scale (Africa, America, Asia and Europe). This choice limits the total number of identities available in the dataset because Wikipedia exhibits a strong bias toward mens’ biographies [35]. A classifier derived from [4], which counts occurrences of the third-person singular feminine (she, her) and masculine (he, him, his) pronouns, is applied. The authors of [4] report 100% accuracy and a verification of 500 identities from FaVCI2D confirms their conclusion.

3.2. Image collection

The Bing Images search engine is used to download images for Wikipedia entries selected as described above. A first download includes up to 150 images per identity without license-related restriction. This image set is used to create a visual representation of identities used only for pairs validation, as described below. A second download collects up to 10 images with a reusable license per identity.

3.3. Image preprocessing and reranking

We aim to create a visual prototype for each identity in order to guide manual pair verification. Face detection is applied to all downloaded images using MTCNN [38]. Then, features are extracted for the detected faces using the ir50 model from [39] and L2-normalized. Third, a mean ID feature is computed from the first 10 images returned by Bing in which only one face was detected. This condition is necessary because it is initially impossible to know which one of the several faces in an image is relevant. We assume that the top Bing results are on average more relevant than the following ones. Finally, all images of the identity are compared with the mean representation. In parallel, a similarity matrix is computed among all the identities of the dataset using the mean representations made of top-10 ranked images. This matrix is exploited to create FaVCI2D variants used in the evaluation. Face features are also extracted for reusable images of the identity and are compared to the mean representations of all candidate identities. Only faces that are closer to the current identity than to any other identity are retained for validation.

3.4. Validation of image pairs

This process is done in two steps. An interface is created to select genuine pairs (see the interface in the supplementary material). A first annotator is instructed to select two of the reusable images which are relevant for the target identity. Feedback is provided in the interface about the difference of age for the pairs which were already selected. Whenever several reusable images are available, the selection of candidates is guided by two related criteria: (1) the faces should be visually different and (2) the age difference should belong to one of the underrepresented bins is favored. Age difference is important insofar the faces change over time, but was not studied previously. A second interface is created for further verification of pairs. Genuine pairs selected during the first step are checked by two more annotators. They are kept only if both agree that the faces represent the same person. The verification of challenging imposters is also done by three annotators to ensure that images of the same ID were not mistakenly kept in the pair.
3.5. FaVCI2D characteristics

The proposed dataset includes identities from 153 countries. 30 countries of the initial 183 were excluded because they are heavily underrepresented. The distribution of genuine identities for different countries is provided in the supplementary material. The total number of unique IDs is 52,411, with 12,468 of them being used in genuine pairs. The total number of images is 64,879, with two images for IDs from genuine pairs and one for the imposter-only IDs. The complete versions of FaVCI2D, created with random and challenging imposter selection, include a total of 24,936 pairs divided equally between the two types of pairs.

We target a balanced gender and geographic distribution. It was possible to obtain enough pairs for America, Asia and Europe but not for Africa. The dataset includes 3,708 genuine pairs, 50% female - 50% male, for each of the first three regions and 1,344 for Africa, 23.3% female - 76.7% male. The gender distribution of IDs in the entire dataset is 44% female - 56% male, which is the closest we can get with reusable resources to a perfect balance.

Age-related information was found for 6,535 out of a total of 12,468 genuine pairs. The distribution of ages at the moment when the photo was taken is: 17% for 18-25 years old; 27% for 26-35; 18% for 36-45; 15% for 46-55; 12% for 56-65; 7% for 66-75 and 4% for 76 and over. The distribution of age difference (in years) between two photos in genuine pairs is 18.5% for the same age, 29.5% for 1 and 2, 23% between 3 and 5; 17% between 6 and 10; and 12 for more than years. While relatively imbalanced the two age-related distributions, they include enough examples in each range to run an age-oriented analysis of verification results.

4. Experimental Validation

Different variants of our dataset are tested depending on the objective of each experiment. The number of genuine and imposter pairs is balanced in all configurations. A thorough evaluation of five state-of-the-art face verification models is proposed. First, these feature extractors are evaluated on two existing datasets. Second, we compare the behavior of these models using challenging/random imposters and a variable size of the pool of imposter IDs. Third, we examine the relation between accuracy and gender. Fourth, we compare the results obtained for 20 countries from four major regions of the world. Finally, we present results obtained for different age ranges and age differences.

4.1. Evaluation with existing datasets

The following models were used in experiments: insightface [9], based on ResNet-150, trained on MS-Celeb1M dataset using ArcFace loss; ir152 [39], based on ResNet-152, trained on MS-Celeb1M dataset using Focal loss; se-gface [17], based on ResNet-27 trained on MS-Celeb1M using the L2-SphereFace loss and fine-tuned on Celeb-Seq dataset; vgg [7], based on SE-ResNet-50 trained on MS-Celeb1M dataset and fine-tuned on VGGFace2 dataset using Softmax loss; facenet [30], based on Inception ResNet, trained on VGGFace2 dataset using SoftMax loss.

In Table 2, we present the results obtained with the five models on LFW [21] and YTF [37]. When available, the original model performance is reported in parenthesis. The results reproduced here are coherent with the original ones. This finding validates the fact that the feature extractors are configured correctly and their further comparison is fair.

4.2. Influence of imposter selection

In Table 3, we present the accuracy of feature extractors in different configurations of imposter pair selection. The similarity between IDs that form imposter pairs is varied between 1 (usage of the most similar imposter ID) to random, the usual verification scenario. The size of the pool of IDs from which imposters are selected is varied between 1,000 and 52,410, the total number of IDs in FaVCI2D. Globally, the best performance is obtained with insightface and the lowest with facenet. The use of challenging pairs reduces performance quite significantly. insightface is the only method whose accuracy is above 90% in the most challenging settings, i.e., most similar imposter ID and largest pool of imposters. The use of challenging pairs allows a better separability compared to a random selection of imposters. When an entire pool of imposters is used, performance with random selection only varies from 97.64% (ir152) to 98.82% (insightface). The corresponding variation for the most challenging setting (Similar = 1) is from 82.61% (facenet) to 95.75% (insightface).

The imposter pool size has virtually no influence for random selection of imposters. This result indicates that very large distractor sets, such as proposed in Megahface [19] or Trillion Pairs [3], are useless in the random configuration. Inversely, the imposter pool size influences performance when similar imposters are used. The fact that performance is reduced between 30,000 and 52,410 imposter IDs indicates that an even larger number of unique IDs would have been useful in FaVCI2D. However, the performance drop tends to reduce when increasing the imposter pool size. Consequently, the dataset provides a usable approximation of the very large-scale performance of the tested models.

These findings confirm that face verification is still an open research problem, especially when challenging imposter pairs are presented to the system. It would be inter-

| Model | Training Data | LFW (99.80) | YTF (99.80) |
|-------|---------------|-------------|-------------|
| insightface | MS-Celeb1M-ArcFace | 99.87 | 97.94 |
| ir152 | MS-Celeb1M | 99.76 (99.80) | 97.80 |
| se-gface | MS-Celeb1M + Celeb-Seq | 99.80 (99.80) | 98.82 (98.80) |
| vgg | MS-Celeb1M + VGGFace2 | 99.40 | 96.78 |
| facenet | VGGFace2 | 99.55 | 95.12 |

Table 2. Accuracy (%) of feature extractors on LFW and YTF.
We run an ablation experiment to estimate the influence of unique IDs count in FAVC12D. We remove 50% and 25% of IDs and test feature extractors with Similar = 1 from 52410 imposters. Five random samplings are used and accuracy is averaged. The obtained results, detailed in the supp. material, are well aligned with those of the full dataset from Table 3. The maximum differences are observed for vgg and reach 0.25% (85.03 for 50% ablation vs 85.28% for the full dataset) and 0.09% (85.19% for 25% ablation vs 85.28%). This indicates that unique IDs count is sufficient for a global evaluation of performance. However, an enrichment of the dataset remains interesting for the evaluation of different demographic segments.

### 4.3. Influence of gender

The results from Table 4 indicate that accuracy is globally lower for female face verification. The performance gap between genders is larger when more similar faces are used as imposters. We note that there is virtually no difference for random imposter selection. Female pairs are recognized marginally better for insface and seqface, while the opposite is true for ir152 and vgg. The gap is largest for vgg and facenet, reflecting gender distribution imbalance from the face recognition datasets used for training the feature extractors. VGGFace2 [7] has stronger gender imbalance compared to MS-CELEB1M [14]. These datasets include fewer female than male identities. Also, male identities have a larger average number of images associated with them. It would be interesting to verify if gender bias subsists for a feature extractor trained with a gender-balanced dataset. This question should be carefully studied by future face verification but is beyond the immediate scope here.

### 4.4. Influence of origin

We present results obtained for countries of origin in Table 5. Mirroring the results from Table 3, the differences between methods are higher for challenging imposters. A more meaningful comparison of feature extractors can be made with challenging imposters. Insface is best for all countries with challenging imposters. The average performance is best for Europe, followed by America, Asia and Africa. These results reflect the structure of the underlying face recognition datasets which are biased toward Europe and North America. A stronger under-representation of some Asian countries seems to occur in VGGFace2 [7] compared to MS-CELEB1M [14], since results for Asia are lower for vgg and facenet, the two VGGFace2-based models. Performance for American countries often sits between that for Europe and those for Asian and African countries. This is interesting insofar American countries include an important mix of populations from other continents.

Within each region, there can be important differences between countries from the same region, even when their inhabitants would be grouped in the same “race” category in other verification datasets, such as RFW [36] or FairFace [18]. This is, for instance, the case for Nigeria and Ghana in Africa or Mexico and Argentina in America and Japan and China in Asia. The use of country also provides interesting insights into which countries are most under-represented in face recognition datasets used to create the feature extractors. For instance, performance is low for Tunisia, Japan and South Korea for all extractors tested.

Globally, the analysis presented in Table 5 comforts our choice to use the country as a proxy for origin rather than race or skin color which were used previously. It also provides more support to the relevance of using challenging imposters instead of random ones in face verification.
Region = Africa
Region = America
Region = Asia
Region = Europe
S. Africa
Nigeria
Egypt
Ghana
Tunisia
US
European

Results for challenging imposter pairs show an inverse correlation between increasing age difference and performance. Similar to age, the most stable results are obtained for insightface and ir152. Larger drops with increasing age difference are observed for vgg and facenet, while seiface sits in the middle. Differences between models are at least in part due to the underlying datasets, with MS-CELEB1M providing stabler discriminatory power across age difference ranges compared to VGGFace2.

5. Conclusions

We first provided a detailed analysis of face verification datasets which highlights their merits and limitations. We gave attention to legal and ethical aspects which are often discussed only marginally. Compliance with such aspects should contribute to better public acceptance of face verification and avoid controversies that led to the withdrawal of datasets such as MS-CELEB1M, MegaFace and DiF.

This analysis led to the introduction of FaVC2D whose objective is to mitigate, to the extent possible, the limitations of existing datasets while preserving much of their qualities. Focus is put on ensuring wide demographic coverage and on including challenging genuine and imposter pairs. Demographic diversity and balance are obtained for most, but not all of the countries. This situation is due to the fact that raw input data are strongly biased toward some regions of the world. However, the demographic spread, balance and level of detail in FaVC2D is better than that of existing face verification datasets.

Finally, we proposed a fine-grained performance analysis with five deep face recognition models. The evaluation shows that model comparison is more meaningful when using challenging imposter pairs. It also provides interesting insights related to steps needed in order to build a fair face verification process. A wide majority of observed biases are actually due to demographically imbalanced training data used to create face recognition models. Beyond its direct use in verification, the proposed dataset could be used during the constitution of future recognition datasets in order to better calibrate them in terms of demographic coverage.

Acknowledgment This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media. It was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.
References

[1] Chalearn looking at people. http://chalearnlap.cv.c.uab.es/dataset/36/description/. Accessed: 2020-11-12. 3, 4
[2] General data protection regulation. https://gdpr.eu/. Accessed: 2020-11-12. 2
[3] Trillion pairs. http://trillionpairs.deepglimt.com/overview. Accessed: 2020-11-12. 1, 3, 6
[4] David Bamman and Noah A Smith. Unsupervised discovery of biographical structure from text. Transactions of the Association for Computational Linguistics, 2:363–376, 2014. 5
[5] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan F Klare, and Anil K Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. IEEE Transactions on Information Forensics and Security, 9(12):2144–2157, 2014. 1, 2
[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. 2018. 2, 3, 4
[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 67–74. IEEE, 2018. 5, 6, 7
[8] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets. Efficient cnns for accurate real-time face verification on mobile devices. In Chinese Conference on Biometric Recognition, pages 428–438. Springer, 2018. 1
[9] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 6
[10] Observatoire des Libertés Numériques. Ban security and surveillance facial recognition, 2019. 2
[11] Coalition for Critical Technology. Abolish the #techtoprisepipeline, 2020. 2
[12] Patrick J Grother, Mei L Ngan, Kayee K Hanaoka, et al. Face recognition vendor test part 3: demographic effects. 2019. 1
[13] International Civil Liberties Monitoring Group. Canadian government must ban use of facial recognition by federal law enforcement, intelligence agencies, 2020. 2
[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In European conference on computer vision, pages 87–102. Springer, 2016. 2, 5, 7
[15] Adam Harvey and Jules LaPlace. Megapixels: Origins and endpoints of datasets created "in the wild", 2019-2020. 2, 4
[16] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. Applied machine learning at facebook: A datacenter infrastructure perspective. In 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), pages 620–629. IEEE, 2018. 1, 2
[17] Wei Hu, Yangyu Huang, Fan Zhang, Ruirui Li, Wei Li, and Guodong Yuan. Seqface: make full use of sequence information for face recognition. arXiv preprint arXiv:1803.06524, 2018. 6
[18] Kimmo Kärkiäinen and Jungsoo Joo. Fairface: Face attribute dataset for balanced race, gender, and age. arXiv preprint arXiv:1908.04913, 2019. 1, 2, 3, 7
[19] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megafake benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4873–4882, 2016. 1, 2, 3, 4, 6
[20] Ruggiero Donida Labati, Angelo Genovese, Enrique Muñoz, Vincenzo Pitrì, Fabio Scotti, and Gianluca Sforza. Biometric recognition in automated border control: a survey. ACM Computing Surveys (CSUR), 49(2):1–39, 2016. 1
[21] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 1, 2, 3, 5, 6
[22] Sandra Soo-Jin Lee, Joanna Mountain, Barbara Koenig, Russ Altman, Melissa Brown, Albert Carrillo, Luca Cavalli-Sforza, Mildred Cho, Jennifer Eberhardt, Marcus Feldman, et al. The ethics of characterizing difference: guiding principles on using racial categories in human genetics. Genome biology, 9(7):404, 2008. 4
[23] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pages 471–478, 2018. 1
[24] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Anil K Jain, W Tyler Niggl, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In 2018 International Conference on Biometrics (ICB), pages 158–165. IEEE, 2018. 1, 2, 3
[25] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. arXiv preprint arXiv:1901.10436, 2019. 1, 2, 3, 4
[26] Emmanuel Pernot-Leplay. China’s approach on data privacy law: A third way between the us and the eu? Penn State Journal of Law & International Affairs, 8(1), 2020. 2
[27] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O’Toole, David Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. The good, the bad, and the ugly face challenge problem. Image and Vision Computing, 30(3):177–185, 2012. 1
[28] Adrian Popescu and Gregory Grefenstette. Spatiotemporal mapping of wikipedia concepts. In Proceedings of the 10th annual joint conference on Digital libraries, pages 129–138, 2010. 5
[29] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–1, 2020. 1, 2, 3
[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015. 6
[31] Ciara Staunton, Santa Slenkenberga, and Deborah Mascalzoni. The gdpr and the research exemption: considerations on the necessary safeguards for research biobanks. European
[32] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *arXiv preprint arXiv:2103.01592*, 2021.

[33] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[34] Richard Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–358, 2020.

[35] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24, 2016.

[36] Mei Wang, Weihong Deng, Jian Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019.

[37] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.

[38] Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtscn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427. IEEE, 2017.

[39] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[40] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, Dec. 2003.
Supplementary Material for "Face Verification with Challenging Imposters and Diversified Demographics"

Adrian Popescu¹, Liviu-Daniel Ştefan², Jérôme Deshayes-Chossart¹, Bogdan Ionescu²
¹Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
²University Politehnica of Bucharest, Romania
{adrian.popescu,jerome.deshayes-chossart}@cea.fr, {liviu.daniel.stefan,bogdan.ionescu}@upb.ro

1. Introduction

In this supplementary material, we provide:

• An illustration of random and challenging imposters.
• Details about the annotation task and the interface used to validate genuine pairs in FaV CI²D.
• The distribution of genuine identities across countries.
• The detailed results obtained after the ablation of a percentage of genuine identities.

2. Illustration of imposters

| Name                  | LFW | FaVCI²D |
|-----------------------|-----|---------|
|                       | Genuine 1 | Genuine 2 | Random imposter | Genuine 1 | Genuine 2 | Similar imposter |
| Menya Moscoso         | ![Image]   | ![Image]   | ![Image]       | ![Image]   | ![Image]   | ![Image]       |
| Akbar Hashemi Rahijani | ![Image]   | ![Image]   | ![Image]       | ![Image]   | ![Image]   | ![Image]       |
| Angela Merkel         | ![Image]   | ![Image]   | ![Image]       | ![Image]   | ![Image]   | ![Image]       |

Figure 1. Illustration of genuine and imposter pairs from LFW and FaVCI²D. LFW pairs are created randomly. Images in FaVCI²D genuine pairs are selected as different from one another among available faces. FaVCI²D imposters are samples of an identity which is visually similar to the genuine identity.

Random and challenging imposters are presented in Figure 1 with examples from LFW and from our dataset proposed in this article, FaVCI²D (Face Verification with Challenging Imposters and Diversified Demographics), respectively. The provided examples illustrate the fact that random imposters make the verification task too easy because the two faces from the random pairs are visually different from one another. The selection of similar imposters provides a more realistic and challenging testbed for face verification. Note also that the annotators are asked to choose challenging imposters whenever there are more than two faces available for an identity.

3. Face annotation task

The annotation of genuine pairs was first performed by one participant. Two versions of the interface were created depending on the availability of age related information for each identity (person). We present the full instructions and interface, which include age-related information, in Figure 2. A similar annotation process was implemented for genuine pairs which did not have age-related information associated to them. In this case, the age-related instructions and associated parts of the interface were naturally omitted.

The participant who started the validation of genuine pairs received the following instructions: "Your participation is needed to create a dataset for face verification, a task whose objective is to automatically detect whether two faces represent the same identity or not. Each page of the interface contained the following elements:

• Up to twelve reference faces which were automatically ranked as representative for the tested identity. These faces are presented in order to facilitate the evaluation of candidate faces. Note that, in some cases, a minority of reference faces might be irrelevant and they should be ignored.

• The birth year of the person which is useful to evaluate if the age associated to candidate faces is realistic.

• At least two candidate faces with associated date taken information. The sources of the date taken information is indicated: EXIF, URL or ALT description. You should select two representative faces which, in addition of being relevant for the identity should: (1) be as different as possible from each other (face orientation, hairstyle, presence of glasses, beard etc.) and (2) have an age difference which is underrepresented in the bins.
to the right. Note that face selection is validated when its background turns to green.

- Age-related statistics capture the number of identities in each age difference bin. The objective is, to the extent possible, to have a balanced representation of these differences.

- The "Save" button allows you to validate a selected pair of faces and move to the next identity.

- The "Ignore" button allows you to skip the current identity if the candidate examples are not representative of the identity. This can be due to: (1) uncertainty about their relevance for the current identity or (2) wrong assignment of the date taken information.”

The essential elements of these instructions were then reminded for each identity. The annotation interface deployed for the annotation of genuine pairs is illustrated in Figure 2. In the example, the two selected faces differ in terms of face orientation and expression. They are also distant in time since the first candidate photo was taken in 2017, while the second was taken in 2009. This age difference also contributes to the creation of a challenging genuine pair of images because faces tend to change over time.

Each genuine pair selected using the interface from Figure 2 is then verified by two more annotators using a simpler interface which displays only the preselected candidate faces and their age when this information is available. The

| Africa          | America        | Europe       |
|-----------------|----------------|--------------|
| South Africa    | 239            | United States| 2100         |
| Nigeria         | 213            | Canada       | 396          |
| Egypt           | 97             | Brazil       | 322          |
| Ghana           | 83             | Argentina    | 267          |
| Tunisia         | 68             | Mexico       | 242          |
| Africa          | 917            | United Kingdom| 576         |
| Japan           | 673            | France       | 353          |
| South Korea     | 540            | Germany      | 341          |
| Philippines     | 216            | Italy        | 290          |
| China           | 209            | Ireland      | 249          |

Table 1. Distribution of the number of genuine pairs for the most represented five countries in each region of the world included in FaVC12D. The total number of genuine pairs is 12468.
two participants were instructed to validate pairs only if they were certain that both faces represented the same identity. The pair is retained only if all three annotators agree that the two selected candidates are relevant.

The selection of most challenging imposters (Similar = 1 in Table 2 of the main paper) should be verified since the two identities can in some rare case be the same. This situation can be an effect of the selection of: (1) fictional characters illustrated by images of the real-life person which is also included in the dataset or (2) from an erroneous selection of reference images for the identity. To make sure that imposter pairs actually contained faces of different identities, we again displayed the pair of faces. Three participants were asked to validate that the two candidate faces belonged to different identities. An imposter pair was retained only if the candidate faces were attributed to different identities by all three participants.

5. Results for ablated dataset

The results from Table 2 complete the ablation study mentioned in Subsection 4.2 of the main paper. We ablate 50% and 25% of the total number of genuine identities from the datasets to verify that the number of unique IDs is sufficient. The results confirm this hypothesis since the variation between the ablated datasets and its full version are small. The maximum difference is obtained for vgg when 50% of IDs are removed and amounts to 0.25%. The differences are under 0.1% for all other models tested in Table 2. Note also that, intuitively, the difference is smaller when the number of removed identities is 25% compared to 50%.

### Table 2. Verification accuracy for the ablation of 50% and 25% of genuine IDs (pairs) from the dataset. Results with the full dataset (0%) are also provided for reference. Results are reported with challenging imposters (Similar=1) selected among 52,410 IDs.

| Model   | Ablation | 50%     | 25%     | 0%     |
|---------|----------|---------|---------|--------|
| insightface |         | 95.69 | 95.72 | 95.75 |
|          | ± 0.05  | ± 0.04 | ± 0.0  |
| ir152   |         | 89.43 | 89.45 | 89.48 |
|          | ± 0.04  | ± 0.04 | ± 0.0  |
| seqface |         | 85.67 | 85.59 | 85.61 |
|          | ± 0.02  | ± 0.02 | ± 0.0  |
| vgg     |         | 85.03 | 85.19 | 85.28 |
|          | ± 0.09  | ± 0.07 | ± 0.0  |
| facenet |         | 82.54 | 82.66 | 82.61 |
|          | ± 0.05  | ± 0.02 | ± 0.0  |

4. Distribution of genuine pairs across countries

We discuss the distribution of genuine pairs a selection of top countries $FaVC12D$. The selected countries correspond those for which results were presented in Table 5 of the main paper. The distribution of identities per country is presented in Table 1. We note that while it was possible to ensure a balanced representation of each large region of the world, imbalance subsists for individual countries. This is mainly an effect of the highly imbalanced character of the frequency of country-related identities in Wikipedia. For instance, nearly a half of initial number of identities with associated photos were from the United States, a percentage which is reduced to less than 17% of genuine pairs kept in $FaVC12D$. While surprising, the relatively low number of pairs from China is explained by the fact that this country has a highly skewed gender distribution in Wikipedia. Nearly 90% of Wikipedia identities with images describe Chinese men. Inversely, it is somewhat surprising to note the relatively large number of Irish identities in the dataset. This situation is probably explained by the fact that Ireland is an English-speaking country and has an active community of Wikipedia contributors.