

How Unique Is a Face: An Investigative Study

Michal Balazia¹, S L Happy¹, François Brémond¹, and Antitza Dantcheva¹

INRIA Sophia Antipolis - Méditerranée, 2004 Route des Lucioles, 06902 Sophia Antipolis, France

{michal.balazia, s-l.happy, francois.bremond, antitza.dantcheva}@inria.fr

Abstract—Face recognition has been widely accepted as a means of identification in applications ranging from border control to security in the banking sector. Surprisingly, while widely accepted, we still lack the understanding of uniqueness or distinctiveness of faces as biometric modality. In this work, we study the impact of factors such as image resolution, feature representation, database size, age and gender on uniqueness denoted by the Kullback-Leibler divergence between genuine and impostor distributions. Towards understanding the impact, we present experimental results on the datasets AT&T, LFW, IMDB-Face, as well as ND-TWINS, with the feature extraction algorithms VGGFace, VGG16, ResNet50, InceptionV3, MobileNet and DenseNet121, that reveal the quantitative impact of the named factors. While these are early results, our findings indicate the need for a better understanding of the concept of biometric uniqueness and its implication on face recognition.

I. INTRODUCTION

Biometrics is the science of identifying humans based on their physical, behavioral or psycho-physiological characteristics [20], with one assumption being that such characteristics are unique. *Uniqueness* refers to the ability of a biometric modality to distinguish between individuals, indicating how a biometric characteristic varies across the population. Consequently, an individual possesses high biometric uniqueness, if the distance distribution of their genuine (mated) comparison scores is well separated from the distance distribution of impostor (non-mated) comparison scores. Hence, an individual is less unique if their genuine and impostor distributions significantly overlap (see Figure 1). We note that while a similarity match score represents a *genuine score* if it is a result of matching two biometric samples of the same individual, an *impostor score* refers to comparing of two biometric samples originating from different individuals [18].

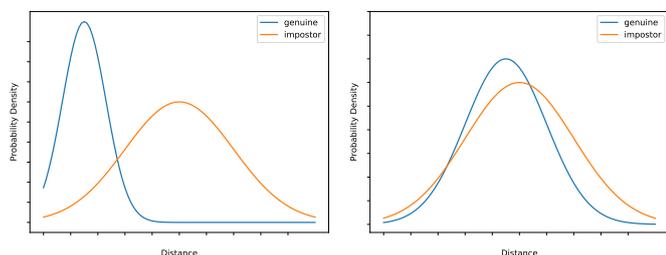


Fig. 1. Genuine and impostor score distributions in a setting with relatively unique subjects (left), as well as a setting with similar subjects (right).

The amount of biometric information is influenced by a set of factors including facial expression, pose, image resolution, distortion, noise or blur [9]. Moreover, the general population

includes family members, a large number of twins (3% in the US between 2014–2018 [25]), as well as doppelgangers¹ (see Figure 2), all of which inherently lower the overall uniqueness of faces in a dataset.

However, the knowledge of distinctiveness of face is incomplete and often relegated to anecdotal interpretation of error rates rather than a systematic exploration of the biology of the characteristics [2], [19], [30]. Hence, we lack an estimate for the upper bound of the amount of discriminatory information contained in a face.

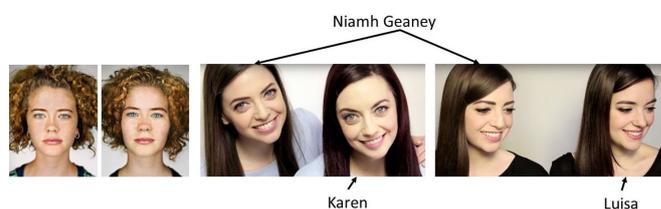


Fig. 2. Are faces unique? Images of identical twins (left) and doppelgangers (right). The anthropology project Twinstrangers has the goal to identify doppelgangers across the world.

Such upper bound can be estimated on the level of extracted features. Generally speaking, class separability of the feature space directly corresponds to accuracy of the estimate. In the context of soft biometrics [6], distinctiveness has to do with collision, or equivalently interference, which describes the event where any two or more subjects belong in the same category of soft biometrics (e.g., female, dark hair, tall) [3]–[5], [7]. We note that this is related to the Birthday paradox [11] and named works answered questions such as: *How large can a population become before it is likelier than not that at least two persons in the group collide biometrically?*

It is known that users of a biometric system may exhibit statistically different degrees of accuracy within the system, which relates to variations in uniqueness across subjects. While some users may experience challenges in authentication, others may be particularly vulnerable to impersonation. The Doddington’s zoo [10], [29], [31] has been frequently used to quantify this phenomenon and specifically to classify users based on verification performance, when users are compared against themselves and against others. Associated major classes include: (a) *sheep*: users who are easy to recognize; (b) *goats*: users who are difficult to recognize; (c) *lambs*: users who are easy to imitate; and (d) *wolves*: users who can easily imitate others. Easy imitation contributes to False Acceptance

¹<https://twinstrangers.net/>

Rate and difficult recognition to False Rejection Rate. This concept was extended in the biometric menagerie [14] (see Figure 3), with additional classes: (e) *chameleons*: users who are easy to recognize and easy to imitate; (f) *phantoms*: users who are difficult to recognize and difficult to imitate; (g) *doves*: users who are easy to recognize and difficult to imitate; and (h) *worms*: users who are difficult to recognize and easy to imitate. Hence, the distribution of a given dataset in such classes impacts the accuracy of a face recognition system on the dataset.

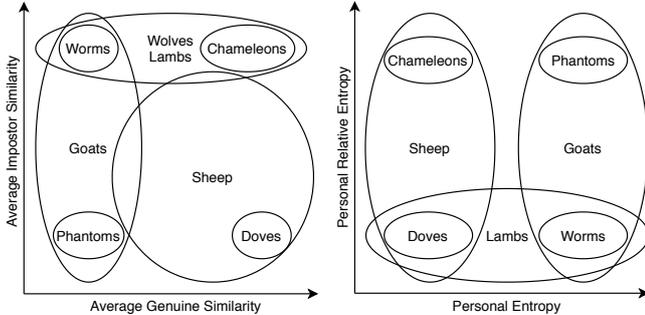


Fig. 3. Animal groups distinguished by Doddington's zoo according to [14]. Left figure shows the distribution for match-based classification and the right one for entropy-based classification.

Motivated by the above, we here aim to provide new insight on the topic of facial uniqueness by investigating the impact of feature extractors (in particular six CNN-based feature extractors: VGGFace, VGG16, ResNet50, InceptionV3, MobileNet and DenseNet121), dataset size (see Section VI-A), image resolution (see Section VI-B), as well as age and gender (see Section VI-C), on the four datasets AT&T, LFW, IMDb and TWINS. Additionally, we identify a potential limitation of the used uniqueness score and propose an additional uniqueness score, which reflects the common occurrence of persons, who look alike (twins, family members and doppelgangers).

II. RELATED WORK

There have been numerous early attempts to quantify biometric uniqueness, distinctiveness and entropy. A key estimator of biometric information related to uniqueness was defined by Adler *et al.* [1] as *the decrease in uncertainty about the identity of a person due to a set of biometric measurements*. This measure was subsequently used by Sutcu *et al.* [36], Takashi and Murakami [38], as well as Gomez-Barrero *et al.* [12]. Similar uniqueness considerations were applied towards improving template protection [12] and performance of biometric systems [21]. Further, facial entropy was quantified and was found to range from 2.6 bits [26] to 55 bits [1]. Recent work of Krivokuća and Marcel [22] on fingervein entropy estimated discriminatory information in fingervein patterns by calculating the number of features used to differentiate between different fingers.

While such works have used different measures of biometric information, what they have in common is the calculation of entropy, which relates to the inter-class variance that can be

artificially increased by including a larger number of features. In order to be independent of the number of extracted features and thus to more accurately estimate the intrinsic information, we also need to consider intra-class variance. Both of these variances are encompassed in our proposed biometric uniqueness score based on the Kullback-Leibler divergence [24], also denoted as *relative entropy*, between genuine and impostor distributions. By its definition below, KL-divergence weighs resistance to both false acceptance and false rejection.

In the context of iris and iris code in particular, an iris uniqueness has been determined by Daugman [8], [9]. Therein 28% of the bits in two IrisCodes are allowed to disagree while still accepting them as a match, resulting in a false match rate of 1 in 92 billion. Given the fixed length of the standardized IrisCode, we consider this rate as an exact indicator of uniqueness in irises.

For the purpose of texture-based generation of fingerprint images, Yankov *et al.* [42] used generative models to estimate upper bounds on the image entropy for systems with small sensor acquisition. They estimated the identification capacity of such systems using the mutual information between different samples from the same finger.

We proceed with describing the uniqueness measure to quantify the identification capacity of faces based on a similar concept.

III. UNIQUENESS AS A DIVERGENCE

Takahashi *et al.* [39] showed that a decrease in uncertainty with respect to identity of an unknown biometric characteristic can be formulated in terms of mutual information

$$I(X; Y) = H(X) - H(X|Y) \quad (1)$$

with $H(X)$ as the marginal entropy, i.e., uncertainty of X , and $H(X|Y)$ as the conditional entropy, i.e., uncertainty of X given the observation of Y . In addition, the authors showed that $I(X; Y)$ can be approximated by the Kullback-Leibler divergence of genuine $P_G(d)$ and impostor $P_I(d)$ probability distributions

$$I(X; Y) \approx D(P_G \| P_I) = \sum_d P_G(d) \log \frac{P_G(d)}{P_I(d)}, \quad (2)$$

where d denote observed dissimilarities between pairs of samples constituting a genuine or impostor observations.

We here note that an average norm estimator can be used to calculate the value of $D(P_G \| P_I)$ from the dissimilarities of samples without computing any probability models [35]. Let G and I be i.i.d. sets of samples with mutual dissimilarities forming the distributions P_G and P_I , respectively. Then, the average norm estimator of the KL-divergence is defined as

$$D(P_G \| P_I) \approx \hat{D}(G, I) = \frac{1}{|G|} \sum_{g \in G} \log \frac{\delta_g(I)}{\delta_g(G)} + \log \frac{|I|}{|G| - 1}, \quad (3)$$

where

$$\delta_s(S) = \frac{1}{|S \setminus \{s\}|} \sum_{s' \in S \setminus \{s\}} \|s - s'\|_2 \quad (4)$$

IV. DATASETS

represents mean of the Euclidean norms between a sample s and all samples from the set S eventually without s .

Calculation of $D(P_G || P_I)$ can be sped up by approximation with random sampling. In order to ensure that the score is not biased towards the distribution of a certain random subset, we calculate the divergence as the average of n different random subset choices. We note that random samples are a sounder representation of the underlying distribution as compared to for example nearest neighbors. Furthermore, for the divergence to mainly depend on the genuine-impostor distributions rather than the count distribution, we choose the subsets $G' \subseteq G$ and $I' \subseteq I$ to contain $r + 1$ and r random samples from G and I , respectively. Equation 3 then transforms to

$$\begin{aligned} \hat{D}_{n,r}(G, I) &= \frac{1}{n} \sum_{k=1}^n \left(\frac{1}{r} \sum_{g \in G'} \log \frac{\delta_g(I')}{\delta_g(G')} + \log \frac{r}{r} \right) \\ &= \frac{1}{nr} \sum_{k=1}^n \sum_{g \in G'} \log \frac{\delta_g(I')}{\delta_g(G')}. \end{aligned} \quad (5)$$

In our experiments, we calculate KL-divergence from the above definition using all available genuine data

$$r = \min(|G| - 1, |I|) \quad (6)$$

so as to obtain the most accurate approximation of the genuine and impostor distributions and

$$n = \lceil 100/r \rceil \quad (7)$$

random subset choices which we believe is sufficiently large to prevent a potential bias. For very large datasets that have too many samples of individual subjects, one can take a smaller r to estimate the divergence for reasons of efficiency. Slightly abusing annotation, in the following we will refer to $\hat{D}_{\lceil 100/\min(|G|-1, |I|) \rceil, \min(|G|-1, |I|)}$ as to simply \hat{D} .

Let a dataset S have c subjects and let $S_p \subset S$ denote a set of samples that belong to subject p . The KL-divergence estimate on this dataset is defined as the average \hat{D} across all subjects

$$\bar{D}(S) = \frac{1}{c} \sum_{p=1}^c \hat{D}(S_p, S \setminus \{S_p\}), \quad (8)$$

where impostor distribution is calculated from the whole remainder of the dataset $S \setminus \{S_p\}$.

Finally, since there is a common practice of normalizing biometric scores to the interval $(0, 1)$, we define the impostor-based biometric uniqueness U of a given dataset S as the estimated and sigmoid-normalized divergence

$$U(S) = \frac{1}{1 + e^{-\bar{D}(S)}}. \quad (9)$$

Note that while $U(S)$ close to 1 indicates that subjects are highly unique, a lower uniqueness indicates that the genuine and impostor distributions are increasingly overlapping. Also note that the uniqueness score is defined as an exponential function with base e , which means that even a small variation in the uniqueness score has a rather significant effect.

In this section we briefly present the four used datasets: AT&T, LFW, IMDB and TWINS. We summarize details in Table I.

- **AT&T Database of Faces (AT&T)** [33] contains face images acquired in the laboratory conditions, with a dark homogeneous background. Subjects are captured in an upright, frontal position. There are 10 different images of each of 40 subjects. Covariates include lighting, facial expressions, open or closed eyes, as well as facial accessories such as glasses.
- **Labeled Faces in the Wild (LFW)** [17] was designed for studying the problem of unconstrained face recognition. The dataset contains 5,749 distinct subjects with 13,233 images of faces collected from the web and hence incorporates highly unconstrained conditions. We select 9,164 images of 1,680 subjects, with two or more images in the dataset. Faces were detected and cropped by the Viola-Jones face detector [40].
- **IMDb-Face (IMDb)** [41] is a large-scale noise-controlled dataset for face recognition research, containing about 1.7M faces of 59K identities, which is a manually cleaned subset of the original 2M raw images. All images were obtained from the IMDb website. We select 1,167,509 images associated to 10,347 identities. IMDb contains gender and age annotations, which allow us to perform additional experiments on subsets pertained to gender and age. See statistics in Table II.
- **ND-TWINS-2009-2010 (TWINS)** [28] dataset was collected by the University of Notre Dame in the years 2009-2010 and comprises 24,050 color photographs of 435 attendee faces captured at the Twins Days Festival under natural light in indoor and outdoor configurations. Facial yaw varies from -90 to $+90$ degrees in steps of 45 degrees. We use a set of cropped 23,762 images, where we detect faces with the MTCNN detector [43].

TABLE I
SIZES OF DATASETS WITH MEAN AND STANDARD DEVIATION STATISTICS, GENDER AND ETHNICITY DISTRIBUTIONS.

	AT&T	LFW	IMDb	TWINS
total number of subjects	40	1,680	10,347	435
total number of samples	400	9,164	1,167,509	23,762
mean/st.d. samples per subject	10.0/0.0	5.5/16.3	112.8/70.5	54.6/36.1
% female/male	10/90	23/77	40/60	75/25
% African/Asian/Caucasian	2/0/98	9/8/83	12/23/65	12/1/84

TABLE II
NUMBERS OF SUBJECTS IN IMDB BY AGE AND GENDER.

full	male	female	0-9	10-19	20-29	30-39
10,347	6,215	4,132	159	856	2558	3031
40-49	50-59	60-69	70-79	80-89	90-99	100-109
1876	1024	522	227	77	14	2

V. ALGORITHMS

We select six state-of-the-art convolutional neural networks (CNNs), which have excelled in a number of face recognition and classification tasks that we proceed to describe.

- **VGGFace** [27] is a 16-layer CNN, trained on the VGGFace dataset comprising over 2M celebrity images. Given a 224×224 input image, the network extracts 4,096 image features from the output of the 6-th fully connected layer.
- **VGG16** [34] constitutes the same 16-layer CNN, however trained on over 1M images from the ImageNet database [32]. We use 4,096 features, provided as output by the second fully connected layer.
- **ResNet50** [13] represents a 50-layer CNN, also trained on ImageNet. As a result of having a large amount of layers, the network has learned rich feature representations for a wide range of images. Similar as the above networks, ResNet50 accepts as input an image of size 224×224 . The 2,048 related ResNet50 features are provided by the last convolutional layer.
- **InceptionV3** [37] is a 48-layer CNN trained on ImageNet. We obtain 2,048 features from the last fully connected layer.
- **MobileNet** [15] is a 56-layer CNN trained on ImageNet, which has been optimized for mobile devices. MobileNet extracts a 1,000-dimensional feature vector.
- **DenseNet121** [16] is a 121-layer CNN trained on ImageNet. The layer structure involves more narrow layers as opposed to ResNet50. The total number of layers is determined by 5 plus two blocks of 58 layers, the last of which results in a 1,024-dimensional feature vector.

VI. EXPERIMENTS

We conduct a set of experiments on the above enlisted datasets with the above named algorithms. We report in each experiment the uniqueness score U , as denoted in Equation 9. We note that small variations in U can indicate large differences in uniqueness due to the exponential function in the denominator in Equation 8. We proceed to investigate U with respect to following factors.

A. Datasets

In the first experiment, we study the uniqueness score across datasets of different size and setting. Related scores are reported in Table III. We note that dataset size is pertinent, as in large datasets it is more likely to encounter similar faces, and the related probability of collision is higher as opposed to small datasets. As expected, AT&T encompasses higher uniqueness scores, due to its small size, as well as constrained acquisition conditions. With respect to features, we see that VGGFace features are systematically outperforming the other networks w.r.t. uniqueness and hence distinctiveness. We believe it is because VGGFace was trained on faces as opposed to other models trained on ImageNet.

TABLE III
UNIQUENESS EVALUATED ON FULL AT&T, LFW, IMDB AND TWINS,
WITH RESOLUTION 224×224 .

	AT&T	LFW	IMDb	TWINS
VGGFace	0.6710	0.5637	0.5420	0.5591
VGG16	0.6417	0.5364	0.5381	0.5339
ResNet50	0.6650	0.5364	0.5330	0.5574
InceptionV3	0.5915	0.5293	0.5272	0.5242
MobileNet	0.6204	0.5353	0.5299	0.5301
DenseNet121	0.6338	0.5302	0.5309	0.5268

B. Image Resolution

We proceed to investigate five image resolutions, namely 224×224 , 112×112 , 64×64 , 48×48 and 36×36 , of the cropped face images of our datasets, all of them in RGB. Given that some of the CNNs take an input of size 224×224 , we apply a lossy data conversion by first downscaling the image to a given resolution and subsequently upscaling it to 224×224 , in order to fit the standardized input size. We observe that resolution has a surprisingly low effect on uniqueness, as shown in Table IV.

For this experiment we provide an additional measure related to uniqueness, namely the image entropy calculated by

$$H = - \sum_{c=1}^C p_c \cdot \log_2(p_c), \quad (10)$$

where C is the color depth, p_c is probability of color c , that is, number of pixels of color c by total number of pixels. We see that while H decreases proportionally with resolution, $U(S)$ is not substantially affected.

TABLE IV
UNIQUENESS (AND MEAN IMAGE ENTROPY) EVALUATED ON FULL AT&T,
LFW, IMDB AND TWINS, WITH VGGFACE FEATURE EXTRACTION
ALGORITHM.

	AT&T	LFW	IMDb	TWINS
224×224	0.6710 (68.8)	0.5637 (87.4)	0.5420 (43.9)	0.5591 (27.8)
112×112	0.6725 (64.0)	0.5636 (70.1)	0.5418 (38.1)	0.5590 (25.3)
64×64	0.6620 (60.4)	0.5635 (59.3)	0.5405 (32.1)	0.5581 (23.3)
48×48	0.6485 (55.5)	0.5637 (34.4)	0.5385 (27.9)	0.5566 (21.9)
36×36	0.6241 (51.2)	0.5603 (19.0)	0.5345 (20.4)	0.5522 (19.7)

C. Subsets

We investigate the effect of *gender* and *age* on the uniqueness score in the dataset IMDb, since related annotations are provided. Results for the gender split are depicted in Table V. Intuitively, the full dataset reflects a higher uniqueness than the gender subsets, which are likely to include more similar faces of only females and only males.

A similar claim can be made w.r.t. age split. Results in Figure 4 show a decrease in uniqueness in all 10-year age-subsets compared to the full IMDb dataset. Moreover, we observe a slight trend of lower uniqueness scores associated with all algorithms for infants and seniors, and of higher uniqueness for middle-aged people. This can be related to the other-age-effect observed in psychology [23].

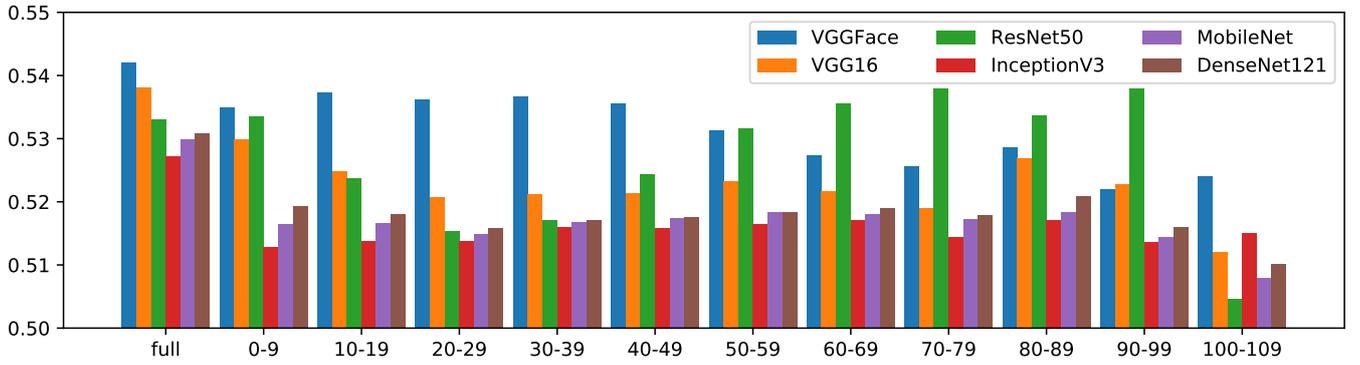


Fig. 4. Uniqueness evaluated on full IMDB compared to its split into 10-year blocks.

TABLE V
UNIQUENESS EVALUATED ON FULL IMDB WITH RESOLUTION 224×224
COMPARED TO ITS SPLITS INTO GENDERS.

	full	female	male
VGGFace	0.542	0.535	0.532
VGG16	0.538	0.520	0.521
ResNet50	0.533	0.519	0.525
InceptionV3	0.527	0.511	0.514
MobileNet	0.530	0.514	0.516
DenseNet121	0.531	0.515	0.517

VII. LIMITATION OF \bar{U} AND INTRODUCTION OF \bar{U}^{MIN}

We recall the definition of the uniqueness score in Equation 8, where impostors are randomly drawn from the whole remainder of the dataset. Under this definition, given a dataset of N subjects, replacing half of the subjects with twins of the already included subjects, would lower the divergence by a factor of $1/2$. Hence the divergence of a dataset of 100 twins will be about 99% of the divergence of non-twin dataset, which is negligible and therefore represents a limitation of the used divergence estimate \bar{D} . For an application that requires a significant impact of occurrence of twins on the uniqueness score, we propose a revised divergence estimate \bar{D}^{MIN} , which places emphasis on the most similar impostor instead of all remaining people. This amendment moves the impostor distribution towards genuine mildly for general population datasets and significantly for twin datasets.

Let a dataset S have c subjects and $S_p \subset S$ denote a set of samples that belong to the subject p . As an alternative, the minimum KL-divergence estimate on this dataset is defined as the minimum \hat{D} across all pairs of subjects

$$\bar{D}^{\text{MIN}}(S) = \frac{1}{c} \sum_{p=1}^c \min_{q \neq p}^c \hat{D}(S_p, S_q), \quad (11)$$

where impostor distribution is calculated from the samples S_q of only the subject q with the closest distribution to S_p . The corresponding alternative uniqueness scores U^{MIN} , calculated using \bar{D}^{MIN} as

$$U^{\text{MIN}}(S) = \frac{1}{1 + e^{-\bar{D}^{\text{MIN}}(S)}}, \quad (12)$$

are reported in Table VI. We observe that while the AT&T dataset based on \bar{D}^{MIN} has a slightly lower $U^{\text{MIN}}(\text{AT\&T})$ score than $U(\text{AT\&T})$, the TWINS dataset drops $\bar{D}^{\text{MIN}}(\text{TWINS}) \approx 0$ and so $U^{\text{MIN}}(\text{TWINS}) \approx 0.5$, as desired.

TABLE VI
UNIQUENESS EVALUATED ON 224×224 RESOLUTION WITH ORIGINAL
AND MINIMUM DIVERGENCE ESTIMATES.

	U		U^{MIN}	
	AT&T	TWINS	AT&T	TWINS
VGGFace	0.671	0.559	0.632	0.507
VGG16	0.642	0.534	0.582	0.491
ResNet50	0.665	0.557	0.542	0.449
InceptionV3	0.592	0.524	0.542	0.485
MobileNet	0.620	0.530	0.570	0.498
DenseNet121	0.634	0.527	0.577	0.494

VIII. CONCLUSIONS

In this work we presented preliminary results on the impact of factors such as image resolution, gender, age, datasets, as well as feature extraction algorithms on facial uniqueness. This is the first work that systematically studies such factors. We provided clear experimental evidence of decrease in the uniqueness score, in the case that (a) image resolution decreases, (b) a single gender is observed, (c) a smaller age group is observed, (d) a larger dataset is used, as well as (e) different feature extractors are used. We illustrated that while feature representation and dataset size significantly affect the uniqueness score, image resolution has a negligible impact. Further, we proposed an alternative uniqueness estimate, which reflects on the presence of twins. Future work will involve establishing a more detailed experimental protocol that among others will aim at quantifying the impact of facial symmetry on uniqueness. Further, common notions of facial entropy, distinctiveness, diversity, complexity, averageness and attractiveness, and their associated relations are to be explored.

ACKNOWLEDGMENT

This work was supported by the French National Research Agency (ANR) under grants ENVISION ANR-17-CE39-0002 and IDEX UCA^{JEDI} ANR-15-IDEX-01.

REFERENCES

- [1] Andy Adler, Richard Youmaran, and Sergey Loyka. Towards a measure of biometric information. In *2006 Canadian Conference on Electrical and Computer Engineering*, pages 210–213. IEEE, 2006.
- [2] National Research Council et al. *Strengthening forensic science in the United States: a path forward*. National Academies Press, 2009.
- [3] A. Dantcheva, J. Dugelay, and P. Elia. Soft biometrics systems: Reliability and asymptotic bounds. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2010.
- [4] A. Dantcheva, A. Singh, P. Elia, and J. Dugelay. Search pruning in video surveillance systems: Efficiency-reliability tradeoff. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1356–1363, 2011.
- [5] Antitza Dantcheva, Jean-Luc Dugelay, and Petros Elia. Person recognition using a bag of facial soft biometrics (BoFSB). In *2010 IEEE International Workshop on Multimedia Signal Processing (MMSp)*, pages 511–516. IEEE, 2010.
- [6] Antitza Dantcheva, Petros Elia, and Arun Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2015.
- [7] Antitza Dantcheva, Carmelo Velardo, Angela D’angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.
- [8] John Daugman. Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935, 2006.
- [9] John Daugman. Information theory and the iriscodes. *IEEE Transactions on Information Forensics and Security (TIFS)*, 11(2):400–409, 2015.
- [10] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, National Inst of Standards and Technology Gaithersburg Md, 1998.
- [11] Philippe Flajolet, Daniele Gardy, and Loÿs Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.
- [12] Marta Gomez-Barrero, Christian Rathgeb, Kiran B Raja, R Raghavendra, and Christoph Busch. Biometric symmetry: Implications on template protection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 941–945. IEEE, 2017.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
- [14] Nesma Houmani and Sonia Garcia-Salicetti. On hunting animals of the biometric menagerie for online signature. *PloS one*, 11(4):e0151691, 2016.
- [15] Andrew Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 04 2017.
- [16] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [18] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.
- [19] Anil K. Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80 – 105, 2016.
- [20] Anil K Jain, Arun A Ross, and Karthik Nandakumar. *Introduction to biometrics*. Springer Science & Business Media, 2011.
- [21] Brendan F Klare and Anil K Jain. Face recognition: Impostor-based measures of uniqueness and quality. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 237–244. IEEE, 2012.
- [22] V. Krivokuca and S. Marcel. Towards quantifying the entropy of fingervin patterns across different feature extractors. In *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8, Jan 2018.
- [23] Dana Kuefner, Viola Macchi Cassia, Marta Picozzi, and Emanuela Bricolo. Do all kids look alike? evidence for an other-age effect in adults. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):811, 2008.
- [24] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.
- [25] Joyce Martin and Michelle Osterman. Is twin childbearing on the decline? twin births in the united states, 2014–2018. *US National Center for Health Statistics Data Brief*, 10 2019.
- [26] Lawrence O’Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, 2003.
- [27] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.
- [28] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt. Distinguishing identical twins by face recognition. In *Face and Gesture 2011*, pages 185–192, March 2011.
- [29] Norman Poh, Samy Bengio, and Arun Ross. Revisiting doddington’s zoo: a systematic method to assess user-dependent variabilities. Technical report, IDIAP, 2006.
- [30] Arun Ross, Sudipta Banerjee, Cunjian Chen, Anurag Chowdhury, Wahid Mirjalili, Renu Sharma, Thomas Swearingen, and Shivangi Yadav. Some research problems in biometrics: The future beckons. In *Int’l Conf. on Biometrics (ICB)*, 2019.
- [31] Arun Ross, Ajita Rattani, and Massimo Tistarelli. Exploiting the “doddington zoo” effect in biometric fusion. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–7. IEEE, 2009.
- [32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Fei Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 09 2014.
- [33] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142, Dec 1994.
- [34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [35] Y. Sutcu, H. T. Sencar, and N. Memon. How to measure biometric information? In *2010 20th International Conference on Pattern Recognition*, pages 1469–1472, Aug 2010.
- [36] Yagiz Sutcu, Elham Tabassi, Husrev T Sencar, and Nasir Memon. What is biometric information and how to measure it? In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, pages 67–72. IEEE, 2013.
- [37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZB Wojna. Rethinking the inception architecture for computer vision. In *arXiv*, 06 2016.
- [38] Kenta Takahashi and Takao Murakami. A metric of information gained through biometric systems. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pages 1184–1187. IEEE Computer Society, 2010.
- [39] Kenta Takahashi and Takao Murakami. A measure of information gained through biometric systems. *Image Vision Comput.*, 32(12):1194–1203, December 2014.
- [40] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.
- [41] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *arXiv preprint arXiv:1807.11649*, 2018.
- [42] M. P. Yankov, M. A. Olsen, M. B. Stegmann, S. S. Christensen, and S. Forchhammer. Fingerprint entropy and identification capacity estimation based on pixel-level generative modelling. *IEEE Transactions on Information Forensics and Security (TIFS)*, 15:56–65, 2020.
- [43] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.