A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Impact of Doppelgängers on Face Recognition: Database and Evaluation

Christian Rathgeb[1], Pawel Drozdowski[1], Marcel Obel[1], André Dörsch[1],
Fabian Stockhardt[1], Nathania E. Haryanto[1], Kevin Bernardo[1], Christoph Busch[1]

Abstract: Lookalikes, a.k.a. doppelgängers, increase the probability of false matches in a facial recognition system, in contrast to random face image pairs selected for non-mated comparison trials. In order to analyse and improve the robustness of automated face recognition, datasets of doppelgänger face image pairs are needed. In this work, we present a new face database consisting of 400 pairs of doppelgänger images. Subsequently, two state-of-the-art face recognition systems are evaluated on said database and other public datasets, including the Disguised Faces in The Wild (DFW) database. It is found that the collected image pairs yield very high similarity scores resulting in a significant increase of false match rates. To facilitate reproducible research and future experiments in this field, the dataset is made available.

Keywords: Biometrics, face recognition, doppelgänger, lookalike, database.

1 Introduction

Face recognition technologies are used in numerous personal, commercial, and governmental identity management systems worldwide. Recent developments in convolutional neural networks have led to remarkable improvements in facial recognition accuracy, surpassing human-level performance [GZ19, Ta14, Ra18]. In particular, state-of-the-art deep recognition systems turn out to be robust against a variety of covariates which may lead to false rejections, such as facial expression [LD20], ageing [BRJ18], or beautification [RDB19].

The improved robustness of said deep face recognition systems may, however, increase the vulnerability against impostors. This has for instance been shown for presentation attacks where an attacker aims at impersonating a target subject by using some attack instrument [MBM18]. In contrast, in a zero-effort impostor attempt, an individual submits their own biometric characteristic while attempting to obtain a successful verification against another subject [IS21]. Previous works reported high success chances for zero-effort impostor attempts in the presence of kin-relationship, in particular for monozygotic, *i.e.* identical, twins [Pr11]. Specific efforts have been devoted to differentiate monozygotic twins in the framework of a facial recognition system, *e.g.* through the analysis of facial marks [Sr12]. It is worth noting that the mentioned effect is far less pronounced for other popular biometric characteristics, *e.g.* fingerprint [Ta12] or iris [DD20].

[1] Hochschule Darmstadt, Germany, contact: christian.rathgeb@h-da.de

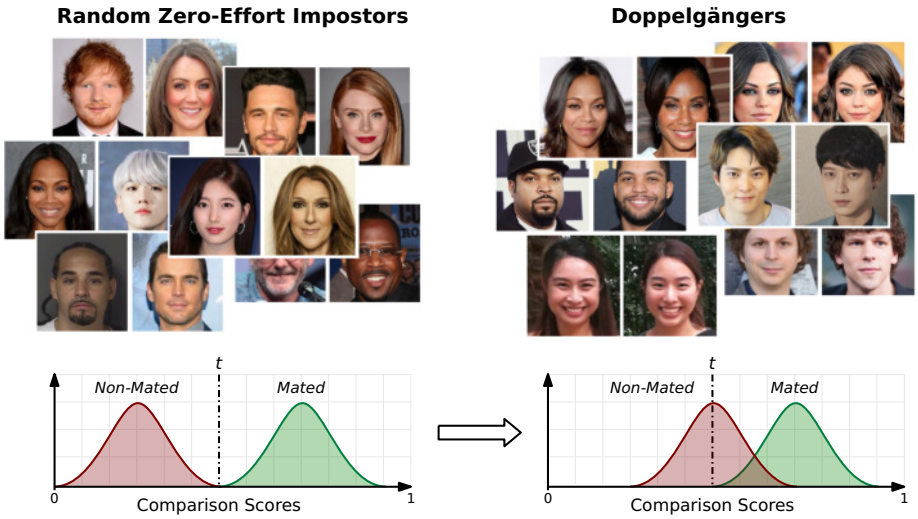


Fig. 1: Random zero-effort impostors (left) achieve low non-mated comparison scores while doppelgängers (right) achieve high non-mated comparison scores and may, if above the decision threshold t, be falsely matched.

In contrast to monozygotic twins, doppelgängers usually refer to biologically unrelated lookalikes. Apart from demographic attributes, doppelgängers also share facial properties such as facial shape. Additionally, some facial properties may further be altered to obtain even higher similarity to a target subject, *e.g.* through the use of makeup [RDB20]. Similar to identical twins, doppelgängers were found to yield high success probabilities compared to random zero-effort impostor attempts, see figure 1. This may lead to serious risks in various scenarios, *e.g.* blacklist checks, where innocent subjects may have a higher chance to match to a lookalike in the list. Lamba *et al.* [La11] presented a preliminary study on the ability of humans and automated face recognition to distinguish lookalikes. Their analysis showed that neither humans nor automatic face recognition algorithms were able to correctly recognise lookalikes. The authors proposed a comparison of facial regions to distinguish lookalikes. Moeini *et al.* [Mo17] suggested to employ 3D reconstruction methods in order to differentiate lookalike faces. To learn highly discriminative facial representations which should also allow to distinguish doppelgängers, Smirnov *et al.* [Sm17] refined the mini-batch selection of a general-purpose face recognition model using a list of lookalikes. Deng *et al.* [De17] introduced the Similar-looking LFW (SLLFW) database, a subset of the Labeled Faces in the Wild (LFW) database, which was selected by human crowdsourcing. It is worth noting that the facial images of LFW are generally unconstrained and of low sample quality. In their Disguised Faces in the Wild (DFW) dataset, Singh *et al.* [Si19] collected facial images which represent challenging face recognition scenarios, including lookalike pairs. More recently, Swearingen and Ross [SR20] presented an approach to improve facial identification performance by re-ranking candidate lists using a lookalike disambiguator which is specifically trained to distinguish between lookalike face images.

Fig. 2: Example doppelgänger image pairs (column-wise) from the collected database.

In this work, we introduce the *HDA Doppelgänger Face Database* consisting of 400 high quality image pairs (with gender parity), which is made publicly available for the research community upon request[3]. Two face recognition systems are evaluated on this newly collected dataset: the well-known open-source ArcFace system and a Commercial-of-the-Shelf (COTS) system. In experiments, the results obtained on the collected dataset are compared with those achieved for lookalikes in the DFW dataset. The rest of this paper is organised as follows: section 2 describes the collected database. Experiments are presented in section 3. Finally, conclusions are given in section 4.

2 Database

The database introduced in this work was collected from the web using search terms like "lookalike" or "doppelgänger". A total number of 400 mostly frontal doppelgänger image pairs was collected and manually checked. During the collection, gender parity as well as diversity in other demographic attributes was assured, resulting in 200 male and female image pairs of various age groups and skin colours. Example image pairs of the collected dataset are shown in figure 2. Similarly to the DFW dataset, the majority of facial images are of celebrities.

3 Experiments

In the experiments, we used the newly collected database described in the previous section as well as a subset of the DFW database [Si19] which contains lookalike face image pairs to investigate the success probability of zero-effort impostor attempts of doppelgängers. In addition to these datasets, mated and non-mated comparison trials were obtained from the FRGCv2 face database [Ph05].

[3] HDA Doppelgänger Face Database:
`https://dasec.h-da.de/research/biometrics/hda-doppelgaenger-face-database/`

Tab. 1: Number of comparisons for the used databases and face recognition systems.

Comparisons	**Ours**		**DFW**	
	ArcFace	COTS	ArcFace	COTS
Doppelgänger	397	389	4,353	4,305
Mated	8,883	6,375	894	893
Non-mated	4,998,147	3,664,320	493,521	496,506

Tab. 2: Descriptive statistics of the used databases and face recognition systems.

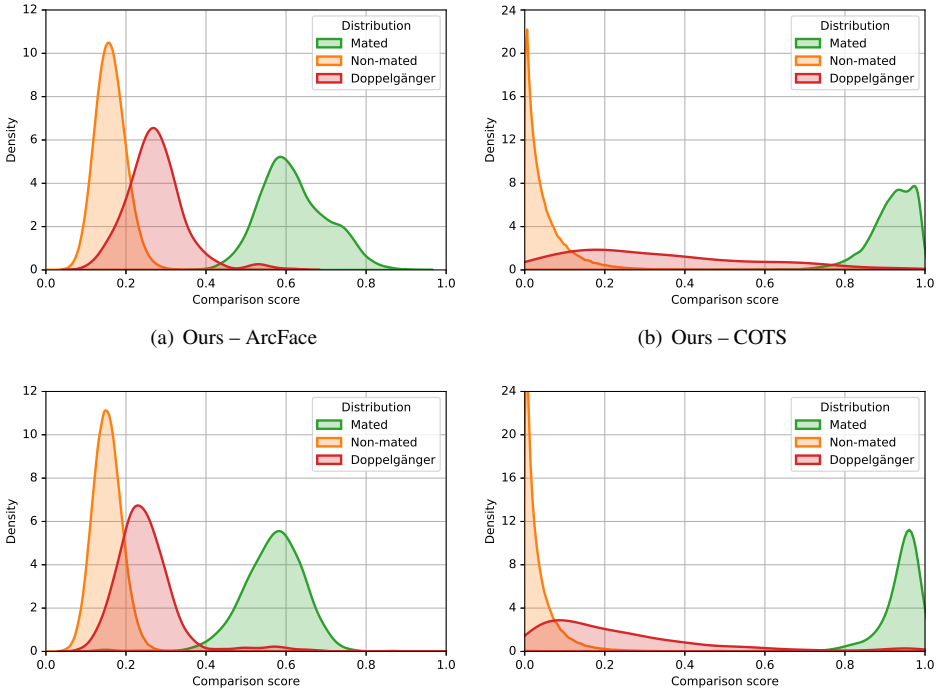
System	**Comparisons**	**Ours**				**DFW**			
		Mean	Std. dev.	Skew.	Ex. kurt.	Mean	Std. dev.	Skew.	Ex. kurt.
ArcFace	Doppelgänger	0.27	0.07	1.12	3.33	0.25	0.08	2.25	9.00
	Mated	0.62	0.08	0.36	-0.23	0.57	0.08	-0.98	3.87
	Non-mated	0.16	0.04	0.36	0.46	0.15	0.04	0.27	0.38
COTS	Doppelgänger	0.34	0.23	0.71	-0.32	0.24	0.22	1.58	2.41
	Mated	0.92	0.06	-1.68	6.13	0.93	0.09	-6.59	57.19
	Non-mated	0.05	0.06	2.63	9.88	0.04	0.05	3.23	17.85

For face recognition, we use a strong open-source system (ArcFace [De19]) with a pre-trained model provided by its authors. ArcFace produces feature vectors of 512 floating-point elements, whose dissimilarity can be computed using the Euclidean distance. For the purposes of visualisation of the results, those dissimilarity scores were mapped into the range $[0, 1]$ using min-max normalisation and converted into similarity scores. While the use of this publicly available and well-known tool facilitates reproducibility, an evaluation with a state-of-the-art commercial off-the-shelf (COTS) system was additionally conducted to increase the practical relevance of the obtained results.

Table 1 summarises the number of comparisons (mated, non-mated doppelgänger, and non-mated) for our dataset and the DFW database for both of the employed face recognition systems. For the COTS system, the number of comparisons tends to be smaller since it failed more often in extracting the face embeddings.

Biometric performance is evaluated using metrics standardised by ISO/IEC [IS21, IS17]. Specifically, biometric recognition performance is reported using false match rate (FMR) and false non-match rate (FNMR); the efficacy of doppelgänger impostor attacks is reported by the impostor attack presentation match rate (IAPMR), *i.e*. the fraction of non-mated doppelgänger comparisons resulting in a false match.

Table 2 lists descriptive statistics of the resulting score distributions which are plotted in figure 3. It can be observed that the comparison scores obtained from lookalike face image pairs are generally higher compared to the non-mated scores. Further, it can be seen that for both face recognition systems, the doppelgängers of the collected dataset tend to yield higher comparison scores than those of the DFW database. Moreover, we observe that doppelgänger score distributions exhibit high standard deviations and longer tails, in



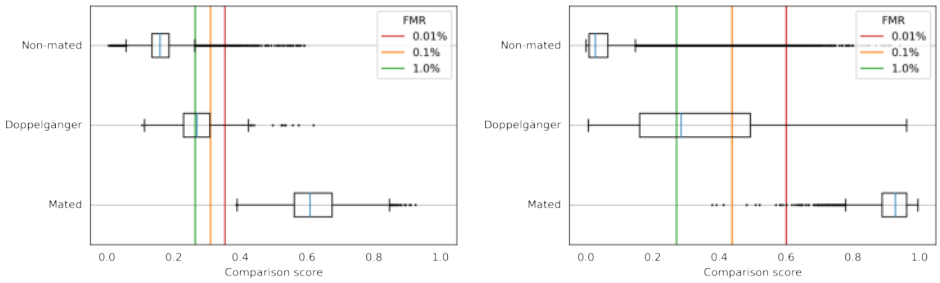
(a) Ours – ArcFace (b) Ours – COTS

(c) DFW – ArcFace (d) DFW – COTS

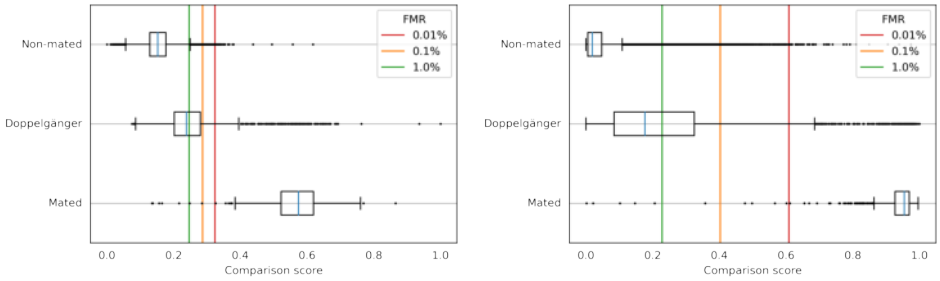
Fig. 3: Probability density functions of scores for both databases and face recognition systems.

particular for the COTS system. That is, some doppelgänger image pairs yield very high comparison scores while the overall distribution is skewed towards the non-mated score distribution. This is further pronounced in the corresponding comparison score boxplots in figure 4 which additionally include decision thresholds obtained from the FMRs. Examples of doppelgängers achieving high comparison scores are shown in figure 5.

Table 3 summarises the performance obtained on both databases in the absence of lookalikes. Here, it can be observed that both face recognition systems obtain competitive recognition performances on both datasets (across the considered, practically relevant [eu15], decision thresholds). The IAPMRs, *i.e.* success chances for doppelgängers, at corresponding decision thresholds are shown in table 4. For a conservative decision threshold, *i.e.* FMR of 0.01%, IAPMRs range from 9.5% to 17% for the collected database for ArcFace and COTS, respectively. As expected based on the analysis of the score distributions, IAPMRs on the DFW database are a bit lower – 6.8% for COTS and 9.6% for ArcFace. For more liberal decision thresholds, *e.g.* FMR of 0.1% or 1%, IAPMRs quickly raise above approximately 25% to 52% for the collected dataset and approximately 17% to 40% on the DFW database. These IAPMR values are alarmingly high and show that the employed face recognition systems are not capable of reliably distinguishing lookalikes. On both



(a) Ours – ArcFace (b) Ours – COTS



(c) DFW – ArcFace (d) DFW – COTS

Fig. 4: Boxplots of scores for both databases and face recognition systems.

Fig. 5: Example doppelgänger image pairs (column-wise) achieving high comparison scores.

datasets, the obtained IAPMR values are significantly higher than the FMRs expected for random non-mated comparisons.

Tab. 3: Performance rates for both databases and face recognition systems.

Database	**System**	**FNMR at FMR of**		
		1.00%	0.10%	0.01%
Ours	ArcFace	0.00%	0.00%	0.00%
	COTS	0.00%	0.05%	0.17%
DFW	ArcFace	0.56%	0.78%	0.78%
	COTS	0.56%	0.67%	1.12%

Tab. 4: Attack success chance of doppelgängers for both databases and face recognition systems.

Database	**System**	**IAPMR at FMR of**		
		1.00%	0.10%	0.01%
Ours	ArcFace	54.16%	24.94%	9.57%
	COTS	52.44%	29.82%	17.22%
DFW	ArcFace	45.26%	21.59%	9.65%
	COTS	39.70%	17.12%	6.85%

4 Conclusion

Many face recognition evaluation protocols randomly pair face images to obtain non-mated comparisons. Obtained non-mated comparison score distribution may then be used to set up decision thresholds at fixed FMRs. It may be concluded that FMRs (and decision thresholds) obtained in such a way overestimate the security of the underlying face recognition system. Furthermore, one may reasonably argue that zero-effort impostor attacks are less likely to be launched by attackers that look very different from the attacked target subject.

The database of doppelgänger image pairs collected in this work allows for a better estimation of face recognition security w.r.t. zero-effort impostor attacks. It was shown, that a large proportion of doppelgängers contained in our dataset falsely results in a match decision for different state-of-the-art face recognition systems. The collected database is made available to the interested researchers upon request. We believe that this may facilitate improvements in face recognition towards a reliable distinction of lookalikes. Further, we would expect that such improvements would enhance the security of face recognition in general as well as against attacks, *e.g.* presentation attacks.

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[BRJ18] Best-Rowden, L.; Jain, A. K.: Longitudinal Study of Automatic Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(1):148–162, 2018.

[DD20] Daugman, J.; Downing, C.: Broken Symmetries, Random Morphogenesis, and Biometric Distance. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(3):271–278, 2020.

[De17] Deng, W.; Hu, J.; Zhang, N.; Chen, B.; Guo, J.: Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership. Pattern Recognition, 66:63–73, 2017.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694, 2019.

[eu15] eu-LISA: Best Practice Technical Guidelines for Automated Border Control ABC Systems. Technical Report TT-02-16-152-EN-N, European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, September 2015.

[GZ19] Guo, G.; Zhang, N.: A survey on deep learning based face recognition. Computer Vision and Image Understanding, 189:102805, 2019.

[IS17] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 30107-3. Information Technology – Biometric presentation attack detection – Part 3: Testing and Reporting, September 2017.

[IS21] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 19795-1:2021. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework, June 2021.

[La11] Lamba, H.; Sarkar, A.; Vatsa, M.; Singh, R.; Noore, A.: Face recognition for look-alikes: A preliminary study. In: International Joint Conference on Biometrics (IJCB). pp. 1–6, 2011.

[LD20] Li, S.; Deng, W.: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, pp. 1–1, 2020.

[MBM18] Mohammadi, A.; Bhattacharjee, S.; Marcel, S.: Deeply vulnerable: A study of the robustness of face recognition to presentation attacks. IET Biometrics, 7(1):15–26, January 2018.

[Mo17] Moeini, A.; Faez, K.; Moeini, H.; Safai, A. M.: Open-set face recognition across look-alike faces in real-world scenarios. Image and Vision Computing, 57:1–14, 2017.

[Ph05] Phillips, P. J.; Flynn, P. J.; Scruggs, T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W.: Overview of the face recognition grand challenge. In: Conference on Computer Vision and Pattern Recognition (CVPR). volume 1. IEEE, pp. 947–954, June 2005.

[Pr11] Pruitt, M. T.; Grant, J. M.; Paone, J. R.; Flynn, P. J.; Bruegge, R. W. Vorder: Facial recognition of identical twins. In: Int'l Joint Conf. on Biometrics (IJCB). pp. 1–8, 2011.

[Ra18] Ranjan, R.; Sankaranarayanan, S.; Bansal, A.; Bodla, N.; Chen, J.; Patel, V. M.; Castillo, C. D.; Chellappa, R.: Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. IEEE Signal Processing Magazine, 35(1):66–83, 2018.

[RDB19] Rathgeb, C.; Dantcheva, A.; Busch, C.: Impact and Detection of Facial Beautification in Face Recognition: An Overview. IEEE Access, 7:152667–152678, October 2019.

[RDB20] Rathgeb, C.; Drozdowski, P.; Busch, C.: Makeup Presentation Attacks: Review and Detection Performance Benchmark. IEEE Access, 8:224958–224973, December 2020.

[Si19] Singh, M.; Singh, R.; Vatsa, M.; Ratha, N. K.; Chellappa, R.: Recognizing Disguised Faces in the Wild. Transactions on Biometrics, Behavior, and Identity Science (TBIOM), 1(2):97–108, March 2019.

[Sm17] Smirnov, E.; Melnikov, A.; Novoselov, S.; Luckyanets, E.; Lavrentyeva, G.: Doppelganger Mining for Face Representation Learning. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 1916–1923, 2017.

[Sr12] Srinivas, N.; Aggarwal, G.; Flynn, P. J.; Vorder Bruegge, R. W.: Analysis of Facial Marks to Distinguish Between Identical Twins. IEEE Transactions on Information Forensics and Security, 7(5):1536–1550, 2012.

[SR20] Swearingen, T.; Ross, A.: Lookalike Disambiguation: Improving Face Identification Performance at Top Ranks. In: 25th International Conference on Pattern Recognition (ICPR. pp. 1–6, 2020.

[Ta12] Tao, X.; Chen, X.; Yang, X.; Tian, J.: Fingerprint Recognition with Identical Twin Fingerprints. PLOS ONE, 7(4):1–7, 04 2012.

[Ta14] Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1701–1708, 2014.