On the Privacy Risks of Large-Scale Processing of Face Imprints

István Fábián, Gábor György Gulyás

Abstract: As technology advances, the number of applications relying on face recognition is on the rise. While facial recognition technologies have many benefits, it's important to use them in a responsible manner in order to avoid privacy risks. In this paper we analyze the privacy risks of the processing of face imprints generated by face recognition technologies. We characterize the risks of re-identification attacks against facial imprint databases in multiple scenarios regarding different attacker strength. Our findings show that even if a large number of subjects are surveilled and the attacker can only inefficiently benefit from using the embeddings, the risk of re-identification is still concerning.

Keywords: facial recognition, privacy, risk, identification, face embedding, deep metrics

1 Introduction

With the trend of technology getting cheaper and the advance of smart technologies, security and surveillance cameras were getting more and more wide spread recently. According to recent news, Chongqing, a single Chinese city alone exhibit 2.58 million surveillance cameras alone [5]. While this is not constrained to distant countries such as China, as London also has 627,707 of such cameras [5]. As the technology is cheap, these devices enable face recognition (hereafter also referred to as FR) technologies at scale. However, we can be certain that using FR widely will have a significant impact on the society as a whole, on the everyday life at companies, and on personal lives as well.

In this paper we consider risks posed by FR. In particular, we look at the case of the large-scale storing and processing of face imprints generated by FR technologies. This technology uses the photo or a video frame containing a person's face to extract an imprint from it. The imprint, or how the literature calls it, the embedding, describes the face based on its unique characteristics. Therefore it can be used for identification. When generated by deep learning techniques, the embedding is usually hard for a human to interpret, as usually it is a vector of real values. Length of the vector may differ, OpenCV [**?**] and DLIB [4] provide embeddings at the length of 128, while others provide longer, e.g., 512 float values [2].

Identification (i.e. the recognition) works by comparing multiple embedding vectors to each other and calculating similarity between them (e.g. via Euclidean distance). At the end, pairwise similarities of the embeddings determine whether the two faces are of the same person. It's presumed that the lower the distance, the higher the similarity, and the similarity of embeddings is proportional to the similarity of the faces. Usually if the distance is below a certain threshold, the embeddings are considered to belong to the same person.

The system setup is the following: cameras observe some areas (for example at a company, or in a public space) and extract facial embeddings of people passing by. Either the cameras themselves are capable of doing the extraction, or they transfer their footage to a capable server device that would do so. Depending on the use case (tracking, authentication, identification, etc.), either embeddings are stored in a database to be used later, or are compared to other embeddings that are already stored in the database.

The reason why storing this may be concerning is that embeddings are biometric data (they are based on features of the human body that the person cannot change) and unlike other biometric data, such as fingerprints, facial images can be easily captured without a person's knowledge and consent, and can be captured in large-scales [1]. Therefore, in this paper, we look at risks related to the processing of embeddings, more specifically we analyze the privacy risk of re-identification by using face imprints.

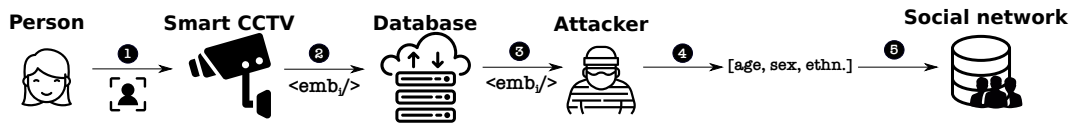


Figure 1: The considered attack when a malicious third party reconstructs demographic data from embeddings and re-identifies the embedding by looking up potential data subjects on social networking sites.

2 Attacker models

Unique pieces of information can be combined indirectly with other data sources in order to put back the names over de-identified data (i.e. where all directly identifying attributes are removed). We call such procedures re-identification attacks. Consider an example where a company publishes a database with information about its employees, de-identified by removing explicitly identifying fields (names, email, etc.) and replacing them with unique random IDs. While this database alone might be considered de-identified, an attacker may link records from this company-related dataset by using the demographic data to the corresponding records in a medical dataset.

In our work we consider re-identification attacks related to embeddings. There are several ways how an attacker can be successful at re-identification by using face embeddings:

- By matching embeddings, e.g. the attacker has a photo, extracts an embedding and looks for a match. As mentioned in the Introduction: if the distance between the two embeddings is below a threshold then the embeddings belong to the same person with some probability.
- If direct search of embeddings is not possible, the attacker could reconstruct the face from the embedding in the database [6], and run a visual search in a face database (e.g. photos on a social network).
- Assuming that embeddings may contain demographic data about the data subject, the attacker can try reconstructing such data from the embedding itself (for example using a machine learning model trained for this task) and looking that up in another database. As reconstruction of faces is possible, this should be possible too, however, the accuracy of such attacks are still questionable (we leave the exploration of this as future work).

In our work, we consider the third class of attacks. Latanya Sweeney showed in 2000 that the zip code, sexuality and date of birth combined together provide a unique identifier for 87% of the population based on US census data. [7]

Therefore, if an attacker accurately predicts such demographic data from embeddings, and knows further pieces of background information such as place of work, she may learn the identity of the anonymized person by looking him on social network sites (e.g. on LinkedIn).

Let us explain this attack as follows (see Figure 1). In the 1st step a camera takes the photo of a person and extracts a face embedding, and in the 2nd step this embedding is stored in a database. In the 3rd step the attacker gets the database of embeddings, where she predicts the related demographic data (age, sexuality, ethnicity) in the 4th step. The final 5th step is when the attacker uses this demographic data to re-identify the person on a social network site.

3 Risk level estimation

We demonstrate the previously discussed attack. We consider different scenarios, based on the number of people in the database and the accuracy of the algorithm that's trying to predict

demographic data from the embeddings. To determine the feasibility and threat level of the attack, we ran simulations on the UCI Machine Learning Repository's Adult Dataset [3]. This dataset contains demographic information (including age, sex and race) for more than 30,000 records of people. Furthermore, these records are not of individuals, but of types of individuals, where the "finalweight" column describes the number of people represented by the given record. Our aim with the simulations is to examine what level of re-identification is theoretically possible for a database of 10, 50, 100 and 300 people randomly sampled from [3], if the accuracy for predicting all age, race and sex vary (60%, 75% and 90%). We assumed a machine learning model that can predict age in 10 year intervals.

As per our attacker model, the database sizes were chosen to be reasonable assumptions for the number of employees of a small or medium sized company. To construct the smaller original databases of size 10, 50, 100 and 300 for the simulation, we randomly sampled the required number of entries from [3] using the values in the *'fnlwgt'* column as weights, which indicate the number of people believed to be represented by a given entry. After sampling, according to the prediction accuracy percentage (90%, 75%, or 60%) the corresponding number of rows were left untouched, while the remaining rows' age attributes were randomly permuted to simulate the inaccurate predictions. This random permutation was then repeated with the same prediction accuracy percentage for the other two attributes, too (sex and ethnicity). This way we ended up with three predicted databases for each smaller original database, where all three attributes were predicted with either 90%, 75%, or 60% accuracy. As the last step, for each database we counted what percentage of data subjects were correctly predicted to fall in an equivalence class of size 1, 2-5, 6-10, 11-20 and 20+. We then repeated this procedure 100 times and averaged out the results.

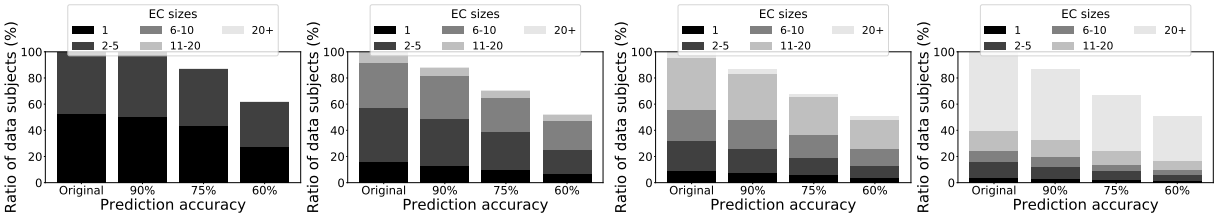
Figures 2a - 2d show our findings. We can observe that there are many records in unique or small equivalence classes both in smaller and larger databases, which pose high privacy risks. The attacker is the most successful at re-identification in the case of the database of 10 with the highest 90% prediction accuracy, when 50.1% of people fall in a unique equivalence class, and all the others fall in an equivalence class of size 2 to 5. If the accuracy is decreased to 60%, still 27.7% falls in a unique equivalence class, and 33.9% falls in an equivalence class of size 2 to 5 (see Figure 2a). Regarding the largest database of 300, 3.75% of people are in a unique equivalence class, and 11.79% are in an equivalence class of size 2 to 5. Even in the worst case scenario for the attacker, which is 60% accuracy for a database of 300, the rate of people in unique equivalence classes doesn't fall below 1.38%, and nor does the rate of people in an equivalence class of size 2 to 5 fall below 4.64% (see Figure 2d). Thus, it's worth noting that while a decrease in accuracy results in a decrease in re-identification probability, risks are not diminished drastically. As a result, due to the privacy risk presented, the actual achievable prediction accuracy must be examined in future work.

In summary, as expected, the smaller the database size, the higher the re-identification risk is, because smaller sized databases have a higher chance of being reconstructed in such a way that people are correctly mapped to an equivalence class of size 1 to 5. Indeed, the higher the prediction accuracy, the higher the re-identification risk is, because the higher percentage of people are predicted to be in the correct equivalence class.

4 Conclusion

In this paper we have presented the potential privacy risks in relation to the processing of face embeddings. We have discussed why face embeddings must be considered sensitive biometric data and we looked at various attacker models that could pose a threat to data subjects' privacy via re-identification.

In particular, we analyzed the risks of re-identification by reverse-engineering demographic data (age, sexuality, ethnicity) from embeddings stored in a database. We found that the smaller



(a) $\mathsf{D} = 10$ (b) $\mathsf{D} = 50$ (c) $\mathsf{D} = 100$ (d) $\mathsf{D} = 300$

Figure 2: The ratio of equivalence classes (EC) in the predicted database (D) for various database sizes and prediction accuracies.

the database and the higher the accuracy of prediction, the higher the re-identification risks are. In the case of a 10 person database and 90% accuracy, 50.1% of people belonged to a unique equivalence class, while this number decreased to 27.7% at 60% prediction accuracy. The risks are also not negligible even for larger databases, because for a database of 300 we showed that at 90% accuracy 3.75% of people are in a unique equivalence class, and 11.79% are in an equivalence class of size 2 to 5.

In the future, we aim to train a machine learning model to find out the actual achievable prediction levels for these demographic data. We are also working on ways to modify the embeddings in such a way to make these predictions harder, without compromising the usability of the embeddings.

Acknowledgments

The research has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013, Thematic Fundamental Research Collaborations Grounding Innovation in Informatics and Infocommunications). Project no. FIEK_16-1-2016-0007 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Centre for Higher Education and Industrial Cooperation - Research infrastructure development (FIEK_16) funding scheme.

Icons made by Pixel perfect, fjstudio, Freepik, Pause08, surang from www.flaticon.com.

References

[1] Daniel Le Métayer Inria Claude Castelluccia. Impact analysis of facial recognition: Towards a rigorous methodology, 2020.

[2] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*, 2019.

[3] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[5] Matthew Keegan. Big brother is watching: Chinese city with 2.6m cameras is world's most heavily surveilled, 2019.

[6] Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates, May 2019.

[7] Latanya Sweeney. Simple demographics often identify people uniquely. 2000.